国家人工智能研发战略计划：2019年更新

一份报告

选择人工智能委员会

的

国家科学技术委员会

2019年6月

每日免费获取报告

1、每日微信群内分享**7+**最新重磅报告；

2、每日分享当日**华尔街日报**、金融时报；

3、每周分享**经济学人**

4、行研报告均为公开版，权利归原作者所有，起点财经仅分发做内部学习。

扫一扫二维码
关注公号
回复：**研究报告**
加入“起点财经”微信群。。

亲爱的同事们，

特朗普总统在2019年2月5日的国情咨文演讲中强调了确保美国在开发构成未来产业的新兴技术（包括人工智能）方面发挥领导作用的重要性。为了反映这一重要性，特朗普总统于2019年2月11日签署了行政命令13859，该命令建立了美国人工智能倡议。该倡议是一种整体政府方法，用于维持美国在人工智能方面的领导地位，确保人工智能使美国人民受益，并反映我们国家的价值观。本行政命令中的第一条指令是联邦机构在其年度预算和规划过程中优先考虑人工智能研究与开发（R&D）。随附的国家AI研发战略计划：2019年更新突出了联邦投资人工智能研发的主要优先事项。

人工智能提供了巨大的机会，导致改善医疗保健，更安全和更有效的运输，个性化教育，重大科学发现，改进制造，提高农业作物产量，更好的天气预报等方面的突破。这些好处主要归功于几十年来联邦对基础人工智能研发的长期投资，这些投资为人工智能系统带来了新的理论和方法，以及允许将人工智能转化为实际应用的应用研究。

由于行业，学术界和非营利组织正在进行大量投资，人工智能研发的前景正变得越来越复杂。此外，人工智能的进步正在迅速发展。因此，联邦政府必须不断重新评估其人工智能研发投资的优先级，以确保投资继续推进该领域的前沿，并且不会不必要地重复行业投资。

在2018年8月，政府指示人工智能专责委员会更新2016年国家人工智能研发战略计划。这一过程始于发布信息请求，以征求公众对如何修订或改进战略的意见。对此RFI的回复以及独立的机构审查将此更新告知了战略计划。

在本战略计划中，确定了八个战略优先事项。前七个策略继续从2016年计划开始，反映了公众和政府的多个受访者重申这些策略的重要性，没有要求删除任何策略。第八项战略是新的，重点关注联邦政府与学术界，工业界，其他非联邦实体和国际盟友之间有效伙伴关系日益增加的重要性，以便在人工智能方面取得技术突破并迅速将这些突破转化为能力。

虽然该计划没有为联邦机构投资确定具体的研究议程，但它确实为联邦人工智能研发投资的整体投资组合提供了期望。这项协调的联邦人工智能研发战略将帮助美国继续在人工智能的前沿领先世界，引领我们的经济，增强国家安全，提高生活质量。

此致

米迦勒·克瑞西俄斯
技术政策总裁助理助理2019年6月21日

目录

执行摘要

人工智能（AI）拥有巨大的希望，几乎可以使社会的各个方面受益，包括经济，医疗保健，安全，法律，运输，甚至技术本身。2019年2月11日，总统签署了行政命令13859，维持美国在人工智能方面的领导地位。[1] 该订单启动了美国人工智能倡议，这是一项促进和保护美国人工智能技术和创新的共同努力。该倡议与私营部门，学术界，公众和志同道合的国际合作伙伴合作实施全政府战略。在其他行动中，该倡议的关键指令要求联邦机构优先考虑人工智能研究与开发（R&D）投资，增强对高质量网络基础设施和数据的访问，确保国家在制定人工智能技术标准方面处于领先地位，并提供为新的人工智能时代准备美国劳动力的教育和培训机会。

为支持美国人工智能倡议，该国家人工智能研发战略计划：2019年更新确定了联邦投资于人工智能研发的优先领域。2019年的更新建立在2016年发布的第一个国家AI研发战略计划的基础上，考虑了过去三年中出现的新研究，技术创新和其他考虑因素。此更新由来自联邦政府的领先AI研究人员和研究管理人员开发，得到了更广泛的民间社会的投入，包括来自美国许多领先的学术研究机构，非营利组织和私营部门技术公司。这些主要利益攸关方的反馈肯定了2016年战略计划各部分的持续相关性，同时也呼吁更加重视人工智能值得信赖，与私营部门合作以及其他必要措施。

国家人工智能研发战略计划：2019年更新为联邦政府资助的人工智能研究制定了一系列目标，确定了以下八个战略重点：

策略1：对人工智能研究进行长期投资。优先考虑对下一代人工智能的投资，这将推动发现和洞察，并使美国成为人工智能的世界领导者。战略2：为人工智能协作开发有效的方法。增加对如何理解
创建有效补充和增强人类能力的AI系统。

策略3：理解并解决人工智能的道德，法律和社会影响。通过技术机制研究人工智能系统，包括道德，法律和社会问题。

策略4：确保AI系统的安全性。了解如何设计可靠，可靠，安全和值得信赖的AI系统。

策略5：为AI培训和测试开发共享的公共数据集和环境。开发并实现对高质量数据集和环境的访问，以及测试和培训资源。

战略6：通过标准和基准测量和评估AI技术。为人工智能开发广泛的评估技术，包括技术标准和基准。

策略7：更好地了解国家AI研发人员的需求。改善研发劳动力发展的机会，从战略上培养一支适合AI的劳动力队伍。

战略8：扩大公私伙伴关系，加速人工智能的发展。与学术界，行业，国际合作伙伴和其他非联邦实体合作，促进人工智能研发持续投资和将进步转化为实际能力的机会。

[1] https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

2019年国家AI研发战略计划简介

人工智能使计算机和其他自动化系统能够执行历史上需要人类认知的任务以及我们通常认为的人类决策能力。在过去的几十年里，人工智能已经取得了巨大的进步，今天承诺提供更好，更准确的医疗服务;加强国家安全;改善运输;更有效的教育，仅举几例。增强的计算能力，大型数据集和流数据的可用性以及机器学习（ML）的算法进步使人工智能开发成为可能，创造了经济的新部门并振兴了行业。随着越来越多的行业采用人工智能的基础技术，该领域将继续在全球范围内推动深刻的经济影响和生活质量改善。

这些进步主要得益于联邦对人工智能研发的投资，美国无与伦比的研发机构的专业知识，以及许多美国最具远见的科技公司和企业家的集体创造力。

2016年，联邦政府发布了第一份国家人工智能研发战略计划，认识到人工智能的巨大希望和持续发展的需要。它的开发是为了指导国家的人工智能研发投资，为改善和利用美国的人工智能能力提供战略框架，并确保这些能力在未来几年为美国人民带来繁荣，安全和提高生活质量。

该计划为投资人工智能的联邦机构确定了几个重点关注的重点领域。这些重点领域或战略包括：对人工智能的持续长期投资;有效的人工智能协作方法;理解和解决人工智能的道德，法律和社会影响;确保AI的安全性;开发用于AI培训和测试的共享公共数据集和环境;通过标准和基准测量和评估AI技术;并更好地了解国家的人工智能研发

2019 更新	**RFI响应通知 2019年国家人工智能研发战略计划**

2018年9月，国家网络和信息技术研究与发展协调办公室发布了信息请求（RFI）[2] 代表人工智能特别委员会，要求所有相关方就2016年国家人工智能研究与发展战略计划提出意见。研究人员，研究组织，专业协会，民间社会组织和个人提交了近50份答复;这些回复可在线获取。[3]

许多回复重申了2016年国家AI研发战略计划中概述的分析，组织和方法。大量答复指出了在制造业和供应链等领域投资人工智能的重要性;卫生保健;医学影像;气象学，水文学，气候学和相关领域;网络安全;教育;数据密集型物理科学，如高能物理学;和运输。自2016年国家AI研发战略计划发布以来，人们对AI技术的转化应用的兴趣肯定有所增加。RFI响应中回应的其他共同主题是开发值得信赖的AI系统的重要性，包括公平，道德，问责制和AI系统的透明度;策划和可访问的数据集;劳动力考虑;以及促进人工智能研发的公私合作伙伴关系。

[2] https://www.nitrd.gov/news/RFI-National-AI-Strategi
[3] https://www.nitrd.gov/nitrdgroups/index.php?title=A

劳动力需求。这项工作具有先见之明：今天，世界各国都纷纷效仿并发布了自己的计划版本。

自国家人工智能研发战略计划制定以来的三年中，新研究，技术创新和现实部署迅速发展。行政当局启动了2019年国家人工智能研发战略计划的更新，以解决这些进步，包括快速发展的国际人工智能环境。

值得注意的是，2019年国家人工智能研发战略计划更新的设计仅涉及解决与推进人工智能技术相关的研究和开发优先事项。它没有描述或推荐与人工智能治理或部署相关的政策或监管行动，尽管人工智能研发肯定会为制定合理的政策和监管框架提供信息。

AI作为管理优先级

自2017年以来，行政当局通过强调其在多个主要政策文件（包括国家安全战略）中对美国未来的作用，阐述了人工智能研发的重要性。[4] 国*防战略，*[5] 和2020财年研发预算优先事项备忘录。[6]

2018年5月，科技政策办公室（OSTP）主办了白宫人工智能美国产业峰会，开始讨论人工智能的承诺以及实现美国人民承诺并保持美国领导地位所需的政策。人工智能的年龄。峰会召集了100多位政府高级官员，来自顶级学术机构的技术专家，工业研究实验室负责人和美国商界领袖。

特朗普总统在2019年2月5日的国情咨文演讲中强调了确保美国在发展构成未来产业的新兴技术（包括人工智能）方面发挥领导作用的重要性。

2019年2月11日，总统签署了行政命令13859，维持美国在人工智能方面的领导地位。[7] 该订单启动了美国人工智能倡议，这是一项促进和保护美国人工智能技术和创新的共同努力。该倡议通过与私营部门，学术界，公众和志同道合的国际合作伙伴的合作和参与实施整体政府战略。在其他行动中，该倡议的关键指令要求联邦机构优先考虑人工智能研发投资，增加对高质量网络基础设施和数据的访问，确保国家在制定人工智能技术标准方面处于领先地位，并提供教育和培训机会。为AI新时代的美国劳动力做好准备。

制定2019年国家人工智能研发战略计划更新

2016年国家人工智能研发战略计划建议，负责推进或采用人工智能的许多联邦机构合作确定关键的研发机会，并支持联邦人工智能研发活动的有效协调，包括校内和校外研究。国家科学技术委员会（NSTC）反映了政府对人工智能的优先顺序，建立了一个新的框架来实施这一建议，由三个独特的NSTC小组组成，这些小组由来自联邦研发机构的成员组成，涵盖（1）高级领导层

[4] https://www.whitehouse.gov/wp-content/uploads/2017/12/NSS-Final-12-18-2017-0905.pdf

[5] https://dod.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf

[6] https://www.whitehouse.gov/wp-content/uploads/2018/07/M-18-22.pdf

[7] https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

和战略愿景，（2）运营规划和战术实施，（3）研究和技术专长。这些小组是：

- 人工智能专题委员会，[8] 由主要负责政府人工智能研发的部门和机构负责人组成，就机构间人工智能研发优先事项向政府提出建议；考虑与工业界和学术界建立联邦伙伴关系；建立结构，以改善政府规划和协调人工智能研发；确定利用联邦数据和计算资源来支持我们的国家AI研发生态系统的机会；并支持技术，国家人工智能劳动力的增长。
- NSTC机器学习和人工智能小组委员会（MLAI）由机构人工智能领导和管理人员组成，是专责委员会的运作和实施部门，负责完成特别委员会的任务；创建和维护国家AI研发战略计划；确定和解决与人工智能研究，测试，标准，教育，实施，外联和相关领域相关的重要政策问题；和相关活动。
- 人工智能研发机构间工作组在NSTC的网络和信息技术研发（NITRD）小组委员会下运作，由来自联邦政府的研究项目经理和技术专家组成，向MLAI小组委员会报告；帮助协调跨机构AI研发计划的工作；作为跨机构AI研发实践社区；并通过NITRD小组委员会年度总统预算补编报告政府范围内的人工智能研发支出。

在2018年9月，专责委员会启动了对2016年战略计划的更新，首先是RFI寻求广泛的社群意见，了解2016年国家AI研发战略计划的七项战略是否以及如何值得修改或更换（见附文）。独立地，执行或资助AI研发的联邦部门和机构进行了自己的评估。

2019年国家人工智能研发战略计划更新概述

人工智能专题委员会，NSTC机器学习和人工智能小组委员会以及NITRD人工智能研发机构间工作组共同审查了有关国家人工智能研发战略计划的意见。2016年计划的最初七个重点领域或战略中的每一个都得到了公众和政府的多个受访者的重申，并没有要求删除任何一项战略。这些战略在2019年战略计划更新中更新，以反映当前的最新技术水平，包括：

战略1：对人工智能研究进行长期投资；

战略2：制定有效的人工智能合作方法；

策略3：理解并解决AI的道德，法律和社会影响；战略4：确保AI系统的安全性；

战略5：为人工智能培训和测试开发共享的公共数据集和环境；战略6：通过标准和基准测量和评估AI技术；和战略7：更好地了解国家AI研发人员的需求。

[8] https://www.whitehouse.gov/wp-content/uploads/2018/05/Summary-Report-of-White-House-AI-Summit.pdf

鉴于私人资助的人工智能研发的快速增长以及工业界对人工智能的快速采用，许多对RFI的回应呼吁加强联邦政府对私营部门的研发投入。因此，2019年更新采用了新的第八个战略：

战略8：扩大公私伙伴关系，加速人工智能的发展。

公众和联邦机构的反馈意见确定了进一步促进人工智能开发和采用的若干具体挑战。这些挑战（其中许多涉及多个机构）提供了更深入的洞察，这些国家AI研发战略计划可以指导美国人工智能研发的过程，并且许多与2019年维护美国领导力执行令的主题密切相关。在人工智能。示例包括以下内容：

- *边*疆研究。尽管机器学习在过去几年中带来了惊人的新功能，但仍需要继续研究以进一步推动ML的前沿，并开发其他方法来应对AI的严峻技术挑战（战略1）。
- *积*极影响。随着人工智能能力的增长，美国必须更加重视开发新的方法，以确保人工智能的影响对未来具有强烈的积极作用（战略1,3和4）。
- *信*任和可解释性。真正值得信赖的人工智能需要可解释的人工智能，特别是随着人工智能系统规模和复杂性的增长这需要人类用户和人类设计者全面了解AI系统（策略1,2,3,4和6）。
- *安*全保障。研究人员必须设计出保持AI系统及其使用的数据安全的方法，以便国家能够利用这项技术提供的机会，同时保持机密性和安全性（战略4,5和6）。
- *技*术标准。随着国家开发技术扩展AI能力和保证，它必须测试和基准测试；当技术准备就绪时，它们应该转变为世界的技术标准（战略6）。
- *劳*动力能力。实现这些目标需要培养目前有限且需求量大的熟练人工智能研发人员队伍；美国必须具有创造力和勇气，在培训和获得领导全球人工智能研究和应用所需的熟练劳动力方面（战略7）。
- *合*作伙伴关系。人工智能研发的进步越来越需要联邦政府与学术界，工业界和其他非联邦实体之间建立有效的伙伴关系，以便在人工智能中产生技术突破，并迅速将这些突破转化为能力（战略8）。
- *与*盟国合作。此外，该计划认识到国际合作对于成功实现这些目标的重要性，同时保护美国AI研发企业免受战略竞争对手和敌对国家的影响。

2019年国家AI研发战略计划更新的结构

此更新的国家AI研发战略计划包含2016版的原始文本，包括以下关于R&D战略的部分（小编辑除外）和前七个战略的原始2016年措辞。对于每项战略，2019年国家研发战略计划的更新在最初七项战略的顶部以阴影框提供；这些突出了战略的更新要求和/或新的重点领域。阴影框下面的文本最初出现在2016年国家AI研发战略计划中，提供了今天仍然重要的观察和背景（请注意，在此期间，一些原始细节可能已经过时）。此外，如前所述，2019年更新中增加了一项新的第八项战略，即扩大人工智能研发中的公共私营伙伴关系。

人工智能研发战略

本AI人力资源研发战略计划中概述的研究重点集中在行业不可能单独解决的领域，因此，最有可能从联邦投资中获益的领域。这些优先事项贯穿整个AI，包括AI感知子领域，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共同需求。由于人工智能的广度，这些优先事项涵盖整个领域，而不是仅关注每个子领域特有的个别研究挑战。为了实施该计划，应制定详细的路线图，以解决与计划一致的能力差距。

战略1中概述的最重要的联邦研究重点之一是持续进行人工智能的长期研究，以推动发现和洞察力。许多投资受到美国联邦政府的高风险，高回报[9] 基础研究已经带来了我们今天依赖的革命性技术进步，包括互联网，GPS，智能手机语音识别，心脏监测器，太阳能电池板，先进电池，癌症治疗等等。人工智能的承诺几乎涉及社会的各个方面，并具有显着的积极的社会和经济效益的潜力。因此，为了在这一领域保持世界领先地位，美国必须将其投资重点放在高优先级的基础和长期人工智能研究上。

许多人工智能技术将与人类一起工作，从而在如何最好地创建以直观和有用的方式与人合作的人工智能系统方面面临重大挑战。[10] 人类和人工智能系统之间的墙壁正逐渐开始腐蚀，人工智能系统增强并增强了人类的能力。如战略2所述，需要进行基础研究，以制定人与人之间互动和协作的有效方法。

人工智能的进步为社会带来了许多积极的好处，并提高了美国的国家竞争力。[11] 然而，与大多数变革性技术一样，人工智能在一些领域存在一些社会风险，从就业和经济到安全，道德和法律问题。因此，随着人工智能科学和技术的发展，联邦政府还必须投资研究，以更好地了解人工智能对所有这些领域的影响，并通过开发符合道德，法律和社会目标的人工智能系统来解决这些问题。 ，如战略3所述。

当前AI技术的一个关键缺口是缺乏确保AI系统安全性和可预测性能的方法。由于这些系统具有不同寻常的复杂性和不断发展的特性，因此确保AI系统的安全性是一项挑战。一些研究重点解决了这一安全挑战。首先，策略4强调需要可由用户信任的可解释和透明的系统，以用户可接受的方式执行，并且可以保证充当用户的意图。人工智能系统的潜在能力和复杂性，加上与人类用户和环境的可能互动，使得投资于提高人工智能技术安全性和控制力的研究至关重要。战略5呼吁联邦政府投资共享公共数据集进行人工智能培训和测试，以推进人工智能研究的进展，并能够更有效地比较替代解决方案。

战略6讨论了标准和基准如何能够集中研发来定义进度，缩小差距，并针对特定问题和挑战推动创新解决方案。标准和基准是

---

[9]“高风险，高回报”的研究指的是具有智力挑战性但有可能对研究领域产生深刻积极的变革性影响的有远见的研究。

[10]见2016年人工智能百年研究报告，重点关注人工智能在2030年的预期用途和影响；https://ai100.stanford.edu/2016-report.

[11]J.弗曼，“这次不同吗？人工智能的机遇与挑战，“经济顾问委员会评论，纽约大学：AI现在研讨会，2016年7月7日。

对于测量和评估AI系统以及确保AI技术满足功能和互操作性的关键目标至关重要。

最后，人工智能技术在社会各个领域日益普及，给人工智能研发专家带来了新的压力。核心人工智能科学家和工程师的机会比比皆是，他们对技术有深刻的理解，可以为推动该领域知识的界限产生新的想法。国家应采取行动，确保有足够的人工智能人才管道。战略7解决了这一挑战。

图1（在2019版本的计划中更新）提供了此AI研发战略计划的整体组织的图形说明。在底部的方框中是横切的，底层基础，影响所有AI系统的发展;战略3-7和新战略8中描述了这些基础。下一层更高（中间一排方框）包括推进AI所需的许多研究领域。策略1-2概述了这些研发领域（包括使用灵感的基础研究）。[12] 图中顶行的方框是预计将受益于AI进步的应用程序示例。人工智能研发战略计划的这些组成部分共同确定了联邦投资的高级框架，可以在该领域产生有影响的进展并带来积极的社会效益。

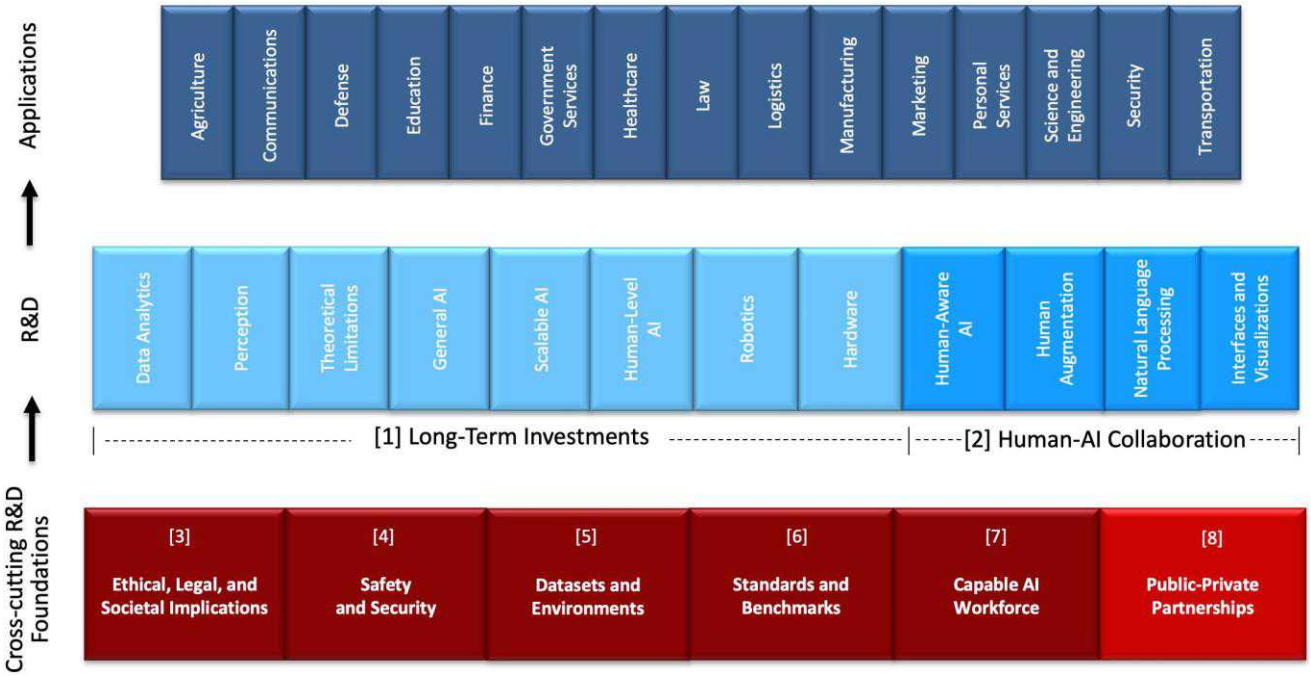

图1. 人工智能研发战略计划的组织（2019年更新，包括战略8）。横切R&D基础的组合（在下排）对于所有AI研究都很重要。许多人工智能研发领域（在中间一行）可以建立在这些横切基础上，以影响各种社会应用（在第一行）。括号内的数字表示该计划中进一步发展每个主题的战略数量。这些策略的排序并不表示重要性。

[12]在整个文件中，“基础研究”包括纯粹的基础研究和基于使用的基础研究 - 所谓的巴斯德象限，由唐纳德斯托克斯在他1997年的同名书中定义 - 指的是用于基础研究的基础研究社会在想。例如，NIH对IT的基本投资通常被称为使用启发式基础研究。

战略1：对人工智能研究进行长期投资

2019更新	持续对基础AI研究进行长期投资

自2016年国家AI研发战略计划发布以来，强大的新功能，主要是ML应用程序，以明确定义的任务，不断涌现。这些能力已经在各种应用中表现出了影响，例如对基因序列进行分类，[20,21] 管理有限的无线频谱资源，[22] 解释医学图像，[23] 和癌症分级。[24] 这些快速进步需要数十年的研究才能使技术和应用成熟。[25] 为了在ML中保持这一进展以实现AI的其他领域的进步，并努力实现通用AI的长期目标，联邦政府必须继续促进ML和AI的长期，基础研究。这项研究将产生转型技术，并反过来在社会各阶层实现突破。

该领域当前的许多进展一直是专门的，定义明确的任务，通常由统计ML驱动，例如分类，识别和回归（即"窄AI系统"）。调查结果

长期，基础的人工智能研究：近期的机构研发计划

自2016年国家航空与战略计划发布以来，一些机构已启动战略1的人工智能研发计划：

- NSF继续资助人工智能的基础研究，包括ML，推理和表示，计算机视觉，计算神经科学，语音和语言，机器人和多智能体系统。NSF与其他机构合作推出了新的联合融资机会 – 尤其是DARPA在高性能，高能效硬件领域的实时性

  以及USDA-NIFA对农业的AI科学[14]– 与行业。[15,16] 此外，NSF正在利用数据革命的大创意[17] 支持对数据科学基础的研究，这将成为未来ML和AI系统的驱动力。
- DARPA于2018年9月宣布对新的和现有的计划进行多年投资，称为"AI Next"活动。[18] 主要活动领域包括提高AI系统的稳健性和可靠性;加强ML / AI技术的安全性和弹性;降低功耗，数据和性能低效;并开创了下一代AI算法和应用，例如可解释性和常识推理。
- *NIH数据科学战略计划*[19] 2018年9月旨在促进生物医学研究界获得数据科学技术和ML / AI能力，以实现数据驱动的医疗保健研究。

[13] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505640&org=NSF
[14] https://www.nsf.gov/pubs/2019/nsf19051/nsf19051.jsp
[15] https://www.nsf.gov/pubs/2019/nsf19018/nsf19018.jsp
[16] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505640
[17] https://www.nsf.gov/cise/harnessingdata/
[18] https://www.darpa.mil/work-with-us/ai-next-campaign
[19] https://datascience.nih.gov/strategicplan
[20] https://ai.googleblog.com/2017/12/deepvariant-highly-accurate-genomes.html
[21] https://irp.nih.gov/catalyst/v26i4/machine-learning
[22] https://www.spectrumcollaborationchallenge.com/
[23] https://news-medical.net/news/20190417/Workshop-explores-the-future-of-artificial-intelligence-in-medical-imaging.aspx
[24] https://www.nature.com/articles/nature21056
[25] https://www.nitrd.gov/rfi/ai/2018/AI-RFI-Response-2018-Yolanda-Gil-AAAI.pdf

该领域已经指出，需要对基础研究进行长期投资，以继续在ML的这些进步基础上再接再厉。此外，需要并行持续努力才能充分实现“通用AI”系统的愿景，该系统在广泛的认知领域中展现出人类智能的灵活性和多功能性。[26,27,28,29]

需要强调开发进一步的ML能力，以交互式和持续学习，感知和注意之间的联系，以及将学习模型纳入综合推理架构。[30] 除了ML之外，人工智能的其他核心领域也需要进行批判性研究，包括常识推理和问题解决，概率推理，组合优化，知识表示，计划和调度，自然语言处理，决策制定和人机交互。这些领域的进步将反过来实现协作机器人以及共享和完全自治的系统（参见策略2）。理解人类智能的巨大挑战需要对共享资源和基础设施进行大量投资.[25] 对ML和AI驱动程序的基础投资也存在广泛的共识，包括数据来源和质量，新颖的软件和硬件范例和平台，以及AI系统的安全性。[31,32] 例如，随着人工智能软件在日常生活和经济的各个方面执行越来越复杂的功能，现有的软件开发范例将需要发展以满足软件生产力，质量和可持续性要求。

最近的联邦投资优先考虑了基础ML和AI研究的这些领域（见边栏），以及ML和AI在众多应用领域的应用，包括国防，安全，能源，交通，卫生，农业和电信。最终，人工智能技术对于解决一系列长期挑战至关重要，例如构建先进的医疗保健系统，强大的智能交通系统以及弹性能源和电信网络。

为了使AI应用程序变得普及，它们必须是可解释和可理解的（参见策略3）。这些挑战对于促进人与人之间的人际关系协作尤为突出（见战略2）。如今，理解和分析AI系统决策并测量其准确性，可靠性和可重复性的能力是有限的。需要持续的研发投资来提高对人工智能系统的信任，以确保满足社会需求并充分满足稳健性，公平性，可解释性和安全性的要求。

对人工智能研发的长期承诺对于继续和扩展当前的技术进步至关重要，并且更广泛地确保人工智能丰富了人类的经验。事实上，2019年“维持美国人工智能领导力行政命令”指出：

> 执行或资助研发（AI研发机构）的执行机构负责人应将AI视为代理机构的研发优先事项，视各自机构的任务而定。这些机构的负责人在制定预算提案和规划时应考虑到这一优先事项。在2020财年和未来几年使用资金。这些机构的负责人还应考虑采取适当的行政行动，以提高对2019年人工智能的关注。

[26] https://ai100.stanford.edu/sites/g/files/sbiybj9861/f/ai100report10032016fnl_singles.pdf
[27] http://cdn.AdQue.org/2018/Ai%20%索引202018%，2020%20RePort.PDF
[28] https://cra.org/ccc/visioning/visioning-activities/2018-activities/artificial-intelligence-roadmap/
[29] https://www.microsoft.com/en-us/research/research-area/artificial-intelligence/
[30] https://cra.org/ccc/events/artificial-intelligence-roadmap-workshop-3-learning-and-robotics/
[31] https://cra.org/ccc/wp-content/uploads/sites/2/2016/04/AI-for-Social-Good-Workshop-Report.pdf
[32] https://openai.com/blog/ai-and-compute/

在具有潜在长期收益的领域，需要进行人工智能研究投资。虽然长期研究的一个重要组成部分是具有可预测结果的渐进式研究，但对高风险研究的长期持续投资可以带来高回报的回报。这些回报可以在5年，10年或更长时间内看到。2012年国家研究委员会的一份报告强调了联邦投资在长期研究中的关键作用，并指出"长期不可预测的潜伏期 – 需要稳定的工作和资金 – 在初始勘探和商业部署之间"。[33] 它进一步指出，"从第一个概念到成功市场的时间通常是在几十年内测量的。"有充分记录的持续基础研究工作的例子包括万维网和深度学习。在这两种情况下，基本的基础始于20世纪60年代;只有经过30多年的持续研究努力，这些想法才能体现到今天在许多类别的人工智能中见证的变革性技术。

以下小节重点介绍其中一些方面。战略2至6讨论了其他类别的重要人工智能研究。

推进以数据为中心的知识发现方法

如*2016年联邦大数据研究与发展战略计划中所述*，[34] 需要许多基本的新工具和技术来实现智能数据理解和知识发现。在开发更先进的机器学习算法时需要进一步的进展，这些算法可以识别隐藏在大数据中的所有有用信息。许多开放式研究问题围绕着数据的创建和使用，包括其对AI系统培训的准确性和适当性。在处理大量数据时，数据的准确性尤其具有挑战性，使人们难以从中评估和提取知识。虽然许多研究通过数据质量保证方法处理数据清理和知识发现的准确性，但还需要进一步研究以提高数据清理技术的效率，创建发现数据不一致和异常的方法，并制定方法。纳入人类反馈。研究人员需要探索新方法，以便同时挖掘数据和相关元数据。

许多AI应用程序本质上是跨学科的，并且使用异构数据。需要进一步研究多模态机器学习以使得能够从各种不同类型的数据（例如，离散的，连续的，文本的，空间的，时间的，时空的，图形的）中发现知识。AI调查人员必须确定培训所需的数据量，并妥善解决大规模和长尾数据需求。除了纯粹的统计方法之外，他们还必须确定如何识别和处理罕见事件;利用知识来源（即解释世界的任何类型的信息，例如重力法或社会规范的知识）以及数据来源，在学习过程中整合模型和本体;当大数据源可能不可用时，用很少的数据获得有效的学习成绩。

增强AI系统的感知能力

感知是智能系统进入世界的窗口。感知始于（可能是分布式的）传感器数据，其具有不同的形式和形式，例如系统本身的状态或有关环境的信息。传感器数据经过处理和融合，通常与先验知识和模型一起，提取与AI系统任务相关的信息，如

[33]国家研究委员会计算机科学电信委员会，信息技术的持续创新（国家科学院出版社，华盛顿特区，2012年），11;https://doi.org/10.17226/13427.

[34] https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf

几何特征，属性，位置和速度。来自感知的综合数据形成态势感知，为人工智能系统提供全面的知识和有效安全地规划和执行任务所必需的世界状况模型。AI系统将极大地受益于硬件和算法的进步，以实现更强大和可靠的感知。传感器必须能够以更高的分辨率和实时捕获更远距离的数据。感知系统需要能够整合来自各种传感器和其他来源（包括计算云）的数据，以确定AI系统当前感知的内容并允许预测未来状态。物体的检测，分类，识别和识别仍然具有挑战性，尤其是在杂乱和动态的条件下。此外，通过使用传感器和算法的适当组合，必须大大提高对人类的感知，以便AI系统可以更有效地与人们一起工作.[10] 需要在整个感知过程中计算和传播不确定性的框架，以量化AI系统在其态势感知中的置信水平并提高准确性。

了解AI的理论能力和局限性

虽然许多AI算法的最终目标是通过类似人类的解决方案解决开放性挑战，但我们并未充分了解AI的理论能力和局限性以及此类人类解决方案甚至可能实现的程度。 AI算法。需要理论工作来更好地理解为什么人工智能技术 - 特别是机器学习 - 在实践中经常运作良好。虽然不同的学科（包括数学，控制科学和计算机科学）正在研究这个问题，但该领域目前缺乏统一的理论模型或框架来理解AI系统的性能。需要对计算可解性进行额外的研究，这是对AI算法在理论上能够解决的问题类别的理解，同样也是对它们无法解决的问题的理解。必须在现有硬件的环境中开发这种理解，以便了解硬件如何影响这些算法的性能。了解哪些问题在理论上无法解决，可以使研究人员针对这些问题制定近似解决方案，甚至开辟新的AI系统硬件研究新路线。例如，当发明于20世纪60年代时，人工神经网络（ANNs）只能用于解决非常简单的问题。在进行并行化等硬件改进后，使用人工神经网络来解决复杂问题变得可行，并且调整算法以利用新硬件。这些发展是实现当今深度学习取得重大进展的关键因素。

开展通用人工智能研究

AI方法可以分为“窄AI”和“一般AI”。窄AI系统在专门的，定义明确的域中执行单独的任务，例如语音识别，图像识别和翻译。几个最近，高度可见，狭窄的AI系统，包括IBM Watson和DeepMind的AlphaGo，已经取得了重大成就。[35,36] 事实上，这些特殊的系统被称为“超人”，因为它们的表现超过了Jeopardy中最优秀的人类玩家！和Go分别。但是这些系统举例说明了狭义的AI，因为它们只能应用于专门设计的任务。在更广泛的问题上使用这些系统需要大量的重新设计工作。相比之下，一般人工智能的长期目标是创建系统

[35]2011年，IBM Watson击败了两名被认为是Jeopardy中最佳人类玩家的玩家！游戏。

[36]2016年，AlphaGo击败了Go，Lee Se-dol的卫冕世界冠军。值得注意的是，AlphaGo结合了深度学习和蒙特卡罗搜索 - 一种在20世纪80年代发展起来的方法 - 它本身建立在20世纪40年代发现的概率方法上。

在广泛的认知领域展示人类智能的灵活性和多功能性，包括学习，语言，感知，推理，创造力和计划。广泛的学习能力将为一般人工智能系统提供将知识从一个领域转移到另一个领域并以交互方式从经验和人类中学习的能力。自人工智能出现以来，人工智能一直是研究人员的野心，但目前的系统还远未实现这一目标。目前正在探索狭义和一般人工智能之间的关系；可以应用一个课程来改进另一个课程，反之亦然。虽然没有普遍的共识，但大多数人工智能研究人员认为，一般人工智能仍然需要几十年的时间，需要长期，持续的研究努力来实现它。

开发可扩展的AI系统

AI系统的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，并且还可以包括与团队一起工作或领导团队的人。这种多AI系统的开发和使用在这些系统的规划，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，以便实时操作并适应环境的变化。它们应该以流畅的方式适应可用通信带宽或系统退化和故障的变化。以前的许多努力都集中在集中规划和协调技术上;但是，这些方法受到单点故障的影响，例如计划员的丢失或与计划员的通信链接的丢失。分布式规划和控制技术在算法上难以实现，并且通常效率低且不完整，但可能对单点故障提供更强的鲁棒性。未来的研究必须发现更有效，更强大，可扩展的技术，用于多个AI系统和人类团队的规划，控制和协作。

促进人类AI的研究

获得类似人类的AI要求系统以人们可以理解的方式解释自己。这将产生新一代智能系统，例如智能辅导系统和智能助理，可有效地协助人们执行任务。然而，当前AI算法的工作方式与人们学习和执行任务的方式之间存在巨大差距。人们只能从几个例子中学习，或通过接受正式指导和/或“提示”来执行任务，或通过观察执行这些任务的其他人。医学院采用这种方法，例如，医学生通过观察一位执行复杂医疗程序的老医生来学习。即使在世界冠军Go游戏这样的高性能任务中，一个大师级玩家也只能玩几千个游戏来训练他/她自己。相比之下，人类需要数百年才能玩出训练AlphaGo所需的游戏数量。关于实现人类AI的新方法的更多基础研究将使这些系统更接近这一目标。

开发更强大，更可靠的机器人

在过去十年中，机器人技术的重大进步正在导致多种应用的潜在影响，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。虽然历史上设想机器人用于静态工业环境，但最近的进展涉及机器人和人类之间的密切合作。现在，机器人技术在补充，增强，增强或模拟人体物理能力或人类智能方面具有前景。然而，科学家需要使这些机器人系统更有能力，更可靠，更易于使用。

研究人员需要更好地了解机器人感知，从各种传感器中提取信息，为机器人提供实时态势感知。在认知和推理方面需要取得进展，以使机器人能够更好地理解物理世界并与之互动。改进的适应和学习能力将使机器人能够概括他们的技能，对他们当前的表现进行自我评估，并从人类教师那里学习一些物理运动。移动性和操纵性是进一步研究的领域，因此机器人可以穿越崎岖不确定的地形并灵活处理各种物体。机器人需要学会以无缝方式团结在一起，并以可靠和可预测的方式与人类合作。

推进硬件以改进AI

尽管人工智能研究通常与软件的进步有关，但人工智能系统的性能在很大程度上取决于其运行的硬件。目前深度机器学习的复兴与基于GPU的硬件技术及其改进的内存的进步直接相关，[37] 输入/输出，时钟速度，并行度和能效。开发针对AI算法优化的硬件将实现比GPU更高的性能水平。一个例子是“神经形态”处理器，其受到大脑组织的松散启发，并且在某些情况下，针对神经网络的操作进行了优化。[38]

硬件的进步还可以提高高度数据密集的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管道的方法。还需要继续研究以允许机器学习算法有效地从高速数据中学习，包括同时从多个数据流水线中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系统智能地对来自大规模模拟，实验仪器和分布式传感器系统（如智能建筑和物联网（IoT））的数据进行采样或优先排序。这些方法可能需要动态I / O决策，其中基于重要性或重要性实时选择存储数据，而不是简单地以固定频率存储数据。

创建AI以改进硬件

虽然改进的硬件可以带来更强大的AI系统，但AI系统也可以提高硬件的性能。[39] 这种互易性将导致硬件性能的进一步提高，因为计算的物理限制需要新颖的硬件设计方法。[40] 基于AI的方法对于改进高性能计算（HPC）系统的运行尤其重要。这种系统消耗大量能量。AI用于预测HPC性能和资源使用情况，并做出可提高效率的在线优化决策;更先进的AI技术可以进一步提高系统性能。AI也可以用来创建

---

[37]GPU代表图形处理单元，它是一个功耗和成本效益的处理器，包含数百个处理内核;这种设计使其特别适用于固有的并行应用，包括大多数AI系统。

[38]神经形态计算是指硬件学习，适应和物理重新配置的能力，从生物学或神经科学中获取灵感。

[39]M.Milano和L. Benini，“HPC系统中工作功耗的预测建模”，“高性能计算会议论文集：31[st]国际会议，ISC高性能2016”（Springer Vol.9697,2016）。

[40]这些计算的物理限制被称为Dennard缩放，并导致高的片上功率密度和称为“暗硅”的现象，其中需要关闭芯片的不同部分以限制温度并确保数据完整性。

自重构HPC系统，可在发生系统故障时处理，无需人工干预。[41]

改进的AI算法可以通过减少处理器和内存之间的数据移动来提高多核系统的性能 – 这是运行速度比当今超级计算机快10倍的亿亿次级计算系统的主要障碍。[42] 实际上，HPC系统中的执行配置永远不会相同，并且不同的应用程序同时执行，每个不同的软件代码的状态在时间上独立发展。人工智能算法需要设计为在线和大规模运行HPC系统。

[41]A. Cocaña-Fernández，J。Ranilla和L. Sánchez，“通过参数学习和混合遗传模糊系统建模在HPC集群中计算节点插槽的节能分配”，“超级计算杂志”71（2015）：1163-1174。

[42]Exascale计算系统每秒至少可以实现10亿次计算。

战略2：开发有效的人工智能协作方法

2019 更新	开发人工智能系统，补充和增强人的能力，更加关注工作的未来

自2016年国家AI研发战略计划发布以来，国家对人工智能合作的兴趣不断增加。当AI系统补充和增强人类能力时，人类和AI成为一系列共享到完全自治场景的合作伙伴。特别是，人工智能合作在工作的未来背景下既是挑战又是机遇。

在过去三年中，新成立的以及长期的会议，研讨会和工作组广泛地优先考虑人与人之间的合作。例如，人类计算和众包会议已经从一个研讨会发展成为一个重要的国际会议，该会议促进了人工智能与人机交互（HCI）交叉的研究。[45] 2018年，人工智能促进协会选择人工智能合作作为其年度会议的新兴主题。[46] 2019年5月，CHI举办了规模最大的人机交互会议，其中包括“弥合人工智能和人机交互之间差距”的研讨会。[47] “人机交互”杂志于2019年3月发出电话，要求就“统一人机交互和人工智能”这一特刊提交意见书。[48]

人工智能协作：最近的代理研发计划

自2016年国家人工智能研发战略计划发布以来，一些机构已开始为战略2做出努力：

- NSF在人类技术前沿的未来工作[43] Big Idea正在支持社会技术研究，使智能技术与人类协同合作，实现对员工的广泛参与，并在一系列工作环境中改善社会，经济和环境效益。
- NOAA（国家海洋和大气管理局）正在推进飓风，龙卷风和其他恶劣天气预报的人工智能合作，其中人类预报员和人工智能系统共同努力改善恶劣天气警报的产生，并确定作为极端前兆的独特模式事件。人类预报员有时被称为“人类在环路之上”，监督人工智能系统的预测并指导结果。
- NIH正在进行自然语言处理研究，该数据库基于从国家医学图书馆维护的所有MEDLINE引文中提取的9630万个事实数据库。
- 2019年美国能源部关于科学机器学习的研讨会报告确定了优先研究方向，主要科学用例以及人类与人工智能合作将改变科学研究方式的新趋势。[44]

[43] https://www.nsf.gov/eng/futureofwork.jsp

[44] *DOE研讨会报告，科学机器学习的基础研究需求：* https://www.osti.gov/biblio/1478744.

[45] 欢迎来到HCOMP 2019：https://www.humancomput

[46] AAAI-18新兴主题人工智能协作：http://www.aaai.org/Conferences/AAAI/2018/aaai18e

[47] 人类在哪里？缩小人工智能和人机交互之间的差距

[48] 呼叫：“统一人机交互和人工智能”人机交互问题：interaction-and-artificial-intelligence-issue-of-human- 计算机交互/。

在工作的背景下，会议开始探索人类，机器及其伙伴关系的作用，如麻省理工学院的计算机科学和人工智能实验室（CSAIL）以及推出年度人工智能和未来的数字经济倡议工作大会。[49,50] 作*为美国人工智能研究20年社区路线图的一部分，*[51] 在2019年，计算社区联盟（CCC）举办了一个研讨会，重点关注人与人工智能系统之间的有意义的互动。[52] 此外，CCC在2017 – 2018年期间运营人类技术前沿工作组，专注于技术的潜力，以提高人员绩效，包括但不限于工作场所，教室和医疗保健系统。[53]

2016年国家人工智能研发战略计划中的交叉战略原则，“人工智能系统的适当信任需要可解释性，特别是随着人工智能在规模和复杂性方面的增长，”在人工智能合作的背景下，已经看到了研发行动呼吁。一些专业协会和机构已将这一原则确定为优先领域（见附文）。该研究领域反映了战略2和3的交叉，因为可解释性，公平性和透明性是人工智能系统与人类有效合作的关键原则。同样，理解和设计人工智能伦理和价值调整系统的挑战仍然是一个开放的研究领域。与此同时，私营部门已经采取了有效的人工智能合作原则。[54,55]

由于联邦机构在过去三年中根据任务目标增加了人工智能投资，因此他们共同强调人机认知，自主权和代理，例如决策支持，风险建模，态势感知和可信机器智能（见边栏）。通过此类研发投资，研究伙伴关系正在越来越多的轴上发展，将计算科学家聚集在一起;行为，认知和心理科学家;和其他领域的科学家和工程师。学术研究人员和工作场所内外人工智能系统用户之间已经形成了新的合作关系。

展望未来，联邦机构必须继续在开放世界中促进人工智能研发，以促进人工智能系统的设计，这些系统包含并适应用户的情况和目标，以便人工智能系统和用户能够在预期和未预料到的情况下协同工作。

虽然完全自主的AI系统在某些应用领域（例如，水下或深空探测）中非常重要，但许多其他应用领域（例如，灾难恢复和医疗诊断）通过人类和AI系统的组合最有效地解决。实现应用目标。这种协作互动利用了人类和AI系统的互补性。虽然人工智能协作的有效方法已经存在，但其中大多数都是“点解决方案”，只能在特定环境中使用特定平台实现特定目标。为每个可能的应用程序实例生成点解决方案无法扩展;因此，需要做更多工作才能超越这些点解决方案

[49] https://futureofwork.csail.mit.edu/.

[50] AI和2019年工作创新峰会的未来：https://analyticsevent.com/.

[51] https://cra.org/ccc/wp-content/uploads/sites/2/2019/03/AI_Roadmap_Exec_Summary-FINAL-.pdf

[52] 人工智能路线图研讨会2 – 互动：https://cra.org/ccc/events/artificial-intelligence- 路线图车间-2-相互作用/。

[53] https://cra.org/ccc/human-technology-frontier/

[54] https://www.microsoft.com/en-us/research/uploads/prod/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf

[55] https://www.partnershiponai.org/about/#our-work

人工智能协作的一般方法。必须在设计适用于所有类型问题的一般系统，需要较少的人力建设和更大的应用程序之间切换设施之间进行权衡，而不是构建大量特定于问题的系统，这些系统可以更有效地解决每个问题。

未来的应用在人类和人工智能系统之间的功能角色划分，人类与人工智能系统之间的相互作用的性质，人类和其他人工智能系统协同工作的数量以及人类和人工智能系统如何沟通和分享态势感知方面会有很大差异. .人与AI系统之间的功能角色划分通常属于以下类别之一：

1. *AI*与人类一起执行功能：AI系统执行支持人类决策者的外围任务。例如，AI可以帮助人们进行工作记忆，短期或长期记忆检索以及预测任务。
2. *当*人类遇到高认知过载时，AI执行功能：AI系统执行复杂的监控功能（例如飞机中的近地警告系统），决策制定以及人类需要帮助时的自动医疗诊断。
3. *AI*代替人类执行功能：AI系统执行人类能力非常有限的任务，例如复杂的数学运算，有争议的操作环境中动态系统的控制指导，有害或有毒环境中自动化系统的控制方面，在系统响应非常迅速的情况下（例如，在核反应堆控制室）。

实现人与人工智能系统之间的有效互动需要额外的研发，以确保系统设计不会导致过度的复杂性，不确定性或推翻。通过培训和经验可以增加人类对人工智能系统的熟悉程度，以确保人类对人工智能系统的能力以及人工智能系统能够做什么和不能做什么有很好的理解。为解决这些问题，应在这些系统的设计和开发中使用某些以人为中心的自动化原则：[56]

1. 采用直观，用户友好的人工智能系统界面，控件和显示设计。
2. 让操作员知情。显示关键信息，AI系统的状态以及这些状态的变化。
3. 让操作员接受培训。参与一般知识，技能和能力（KSA）的复训，以及AI系统采用的算法和逻辑培训以及系统的预期故障模式。
4. 使自动化灵活。部署AI系统应被视为希望决定是否要使用它们的操作员的设计选项。同样重要的是自适应AI系统的设计和部署，可用于在过度工作负荷或疲劳期间为人类操作员提供支持。[57,58]

在创建与人类有效合作的系统时，研究人员面临许多基本挑战。以下小节概述了其中一些重要挑战。

---

[56] C. Wickens和JG Hollands，“注意力，分时和工作量。”在工程，心理学和人类表现（伦敦：Pearson PLC，1999），439-479。

[57] https://www.nasa.gov/mission_pages/SOFIA/index.html

[58] https://cloud1.arc.nasa.gov/intex-na/

寻求人类感知AI的新算法

多年来，AI算法已经能够解决日益复杂的问题。然而，这些算法的能力与人类对这些系统的可用性之间存在差距。需要人性化智能系统，可以直观地与用户进行交互，实现无缝的机器人与人之间的协作。直观的交互包括浅层交互，例如当用户丢弃系统推荐的选项时；基于模型的方法，考虑到用户过去的行为；甚至是基于准确的人类认知模型的用户意图的深度模型。必须开发中断模型，允许智能系统仅在必要和适当时中断人类。智能系统还应具有增强人类认知的能力，知道在用户需要时检索哪些信息，即使他们没有明确地提示系统获取该信息。未来的智能系统必须能够考虑人类的社会规范并采取相应的行动。智能系统如果具有某种程度的情商，就能更有效地与人类合作，以便他们能够识别用户的情绪并做出适当的反应。另一个研究目标是超越一个人和一个机器的交互，转向“系统系统”，即由多个机器与多个人交互组成的团队。

人工智能系统交互具有广泛的目标。人工智能系统需要能够代表多个目标，他们可以采取的行动来实现这些目标，对这些行动的约束以及其他因素，以及轻松适应目标的修改。此外，人类和人工智能系统必须有共同的目标，并且能够相互理解它们以及它们当前状态的相关方面。需要进一步研究以概括人类AI系统的这些方面，以开发需要较少人工工程的系统。

开发用于人体增强的AI技术

虽然人工智能研究的大部分重点都放在与执行狭隘任务的人匹配或表现优异的算法上，但需要开展更多工作来开发能够跨越多个领域增强人类能力的系统。人体增强研究包括在固定设备（例如计算机）上工作的算法；可穿戴设备（如智能眼镜）；植入设备（如大脑界面）；在特定的用户环境中（例如特别定制的手术室）。例如，基于从多个设备组合的数据读数，增强的人类意识可以使医疗助理指出医疗过程中的错误。其他系统可以通过帮助用户回忆过去适用于用户当前情况的经验来增强人类认知。

人类与人工智能系统之间的另一种协作方式是积极学习智能数据。在主动学习中，从领域专家寻求输入，并且仅在学习算法不确定时才对数据进行学习。这是减少首先需要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意见和增加学习算法信任的关键方法。到目前为止，主动学习仅用于监督学习；需要进一步研究将主动学习纳入无监督学习（例如，聚类，异常检测）和强化学习。[59] 概率网络允许以先验概率分布的形式包括领域知识。无论是以数学模型，文本还是其他形式，都必须寻求允许机器学习算法合并领域知识的一般方法。

[59]虽然有监督的学习需要人类提供真实的答案，强化学习和无监督学习却没有。

开发可视化和人工智能接口技术

更好的可视化和用户界面是需要更多开发的其他领域，以帮助人们理解来自各种来源的大量现代数据集和信息。可视化和用户界面必须以人类可理解的方式清楚地呈现日益复杂的数据和从中获取的信息。提供实时结果对于安全关键操作非常重要，并且可以通过增加计算能力和连接系统来实现。在这些类型的情况下，用户需要可视化和用户界面，可以快速传达正确的信息以进行实时响应。

人工智能协作可以应用于各种环境，并且存在通信限制。在某些领域，人工智能通信延迟很低，通信快速可靠。在其他领域（例如，美国国家航空航天局部署漫游者精神和机遇到火星），人类与人工智能系统之间的远程通信具有非常高的潜伏期（例如，地球和火星之间的往返时间为5-20分钟），因此需要部署的平台在很大程度上自主运行，只有高层战略目标传达给平台。这些通信要求和约束是用户界面研发的重要考虑因素。

开发更有效的语言处理系统

人们通过口头和书面语言与人工智能系统进行交互一直是人工智能研究人员的目标。虽然取得了重大进展，但在人类可以与人工智能系统进行有效沟通之前，语言处理必须解决相当大的开放性研究挑战，就像对待其他人类一样。语言处理的最新进展被认为是数据驱动的机器学习方法的使用，这导致了成功的系统，例如，成功地在安静的环境中实时识别流利的英语语音。然而，这些成就只是实现长期目标的第一步。当前的系统无法应对现实世界的挑战，例如嘈杂环境中的语音，重音口音，儿童语音，语音障碍和手语语音。还需要开发能够与人类进行实时对话的语言处理系统。这样的系统需要推断其人类对话者的目标和意图，使用适当的登记，风格和修辞来处理这种情况，并在对话误解的情况下采用修复策略。需要进一步研究开发更容易概括不同语言的系统。此外，需要进一步研究以语言处理系统易于访问的形式获取有用的结构化领域知识。

还需要在许多其他领域中进行语言处理，以使人与AI系统之间的交互更加自然和直观。必须为口头和书面语言的模式构建强大的计算模型，这些模式为情绪状态，情感和立场提供证据，并用于确定言语和文本中隐含的信息。对于在物理世界中运行的AI系统，例如在机器人技术中，需要新的语言处理技术来为环境环境中的语言奠定基础。最后，由于人们在在线交互中进行交流的方式可能与语音交互完全不同，因此必须完善在这些环境中使用的语言模型，以便社交AI系统可以更有效地与人交互。

策略3：理解并解决人工智能的道德，法律和社会影响

2019 更新	解决人工智能中的道德，法律和社会问题

自2016年国家人工智能研发战略计划发布以来，针对人工智能系统开发和部署的道德，法律和社会影响的研发活动有所增加。人们越来越认识到，人工智能系统必须"值得信赖"，人工智能可以改变社会和经济生活的许多领域，包括就业，医疗保健和制造业。经济合作与发展组织（OECD）等国际组织[63] 和G7创新部长[64] 鼓励研发增加对人工智能的信任和采用。

2016年国家AI研发战略计划在确定隐私研究主题方面具有先见之明;通过设计提高AI系统的公平性，透明度和责任性;并设计道德AI的架构。致力于ML和AI系统的公平性，问责制和透明度的研究会议蓬勃发展。[65] 联邦机构已经回应了各种新的研究计划和会议，重点关注这些关键领域（见附文）。

可解释性，公平性和透明度：最近的代理机构研发计划

自2016年国家AI研发战略计划发布以来，一些机构已启动战略3的AI研发计划：

- DARPA的可解释AI（XAI）计划[60] 旨在创建一套ML技术，生成更多可解释的AI系统，同时保持高水平的学习性能（预测准确性）。XAI还将使人类用户能够理解，适当地信任并有效地管理新一代AI系统。更一般地，国防部致力于"引领军事道德和人工智能安全"，作为指导其加速采用人工智能系统的战略方法中列出的五项关键行动之一。[61]
- NSF和亚马逊正在合作[62] 共同支持以人工智能公平为重点的研究，目标是为可信赖的人工智能系统做出贡献，这些系统易于接受和部署，以应对社会面临的重大挑战。感兴趣的具体主题包括但不限于透明度，可解释性，问责制，潜在的不利偏见和影响，缓解策略，公平性验证和包容性考虑。

[60] https://www.darpa.mil/program/explainable-artificia
[61] "2018年国防部人工智能战略总结"：https://media.
1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.
[62] https://www.nsf.gov/pubs/2019/nsf19571/nsf19571.htm
[63] "经合组织关于人工智能的倡议"：http://www.oecd.org/going-digital/ai/oecd-initiatives-on-ai.htm.
[64] "G7创新部长关于人工智能的声明"：http://www.g8.utoronto.ca/employment/2018-labour-annex-b-en.html.
[65] http://www.fatml.org/; https://fatconference.org/; http://www.aies-conference.com/

2019年维持美国人工智能领导力行政命令强调，保持美国在人工智能方面的领导力需要共同努力，以促进技术和创新的进步，同时保护公民自由，隐私和美国价值观：[1]

> 美国必须培养公众对人工智能技术的信任和信心，并在其应用中保护公民自由，隐私和美国价值观，以便充分发挥人工智能技术对美国人民的潜力。

需要更多的研发来开发人工智能架构，通过透明度和可解释性等技术机制将道德，法律和社会问题纳入其中。这项研发将需要技术专家以及其他领域的利益相关者和专家之间的密切合作，包括社会和行为科学，法律，道德和哲学。由于道德决策也可能严重依赖于上下文或应用程序，因此也可能需要与领域专家进行协作。这种跨学科方法可以纳入AI的培训，设计，测试，评估和实施，以便理解和解释AI引起的决策和行动，并减轻意外后果。

因此，联邦机构应通过赞助研究和召集专家和利益相关者，继续促进不断增长的社区对这些问题的进一步研发的兴趣。

当AI代理人自主行动时，我们希望他们的行为符合我们对人类同胞的正式和非正式规范。因此，作为基本的社会秩序力量，法律和道德既可以通知和判断人工智能系统的行为。主要的研究需求包括理解AI的伦理，法律和社会影响，以及开发符合道德，法律和社会原则的AI设计方法。还必须考虑隐私问题;关于这个问题的进一步信息可以在国家隐私研究战略中找到。[66]

与任何技术一样，人工智能的可接受用途将以法律和道德的原则为依据;挑战在于如何将这些原则应用于这项新技术，特别是那些涉及自治，代理和控制的技术。

如“强大和有益的人工智能的研究重点”所阐明的那样。[67]

> 为了构建运行良好的系统，我们当然需要确定每个应用程序域中的良好行为含义。这种道德维度与可用的工程技术，这些技术的可靠性以及进行权衡的问题密切相关 – 计算机科学，机器学习和更广泛的AI专业知识的所有领域都是有价值的。

这一领域的研究可以受益于多学科的观点，涉及计算机科学，社会和行为科学，伦理学，生物医学，心理学，经济学，法律和政策研究专家。需要在NITRD相关IT领域内外的领域（即信息技术以及前面提到的学科）进行进一步调查，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的关键信息技术研究挑战。

---

[66] https://www.nitrd.gov/pubs/NationalPrivacyResearchStrategy.pdf

[67] “公开信：强有力和有益的人工智能的研究重点”（生命未来研究所）：http://futureoflife.org/ai-open-letter/.

通过设计提高公平性，透明度和问责制

人们对数据密集型人工智能算法对错误和误用的敏感性以及性别，年龄，种族或经济阶层可能产生的后果表示了许多担忧。在这方面，适当收集和使用人工智能系统的数据是一项重要挑战。然而，除了纯粹与数据相关的问题之外，人工智能的设计出现了更大的问题，这些问题本身就是公正，公平，透明和负责任的。研究人员必须学习如何设计这些系统，以便他们的行动和决策是透明的，并且可以被人类轻易解释，因此可以检查他们可能包含的任何偏见，而不仅仅是学习和重复这些偏见。关于如何表达和"编码"价值和信仰系统存在严重的知识问题。科学家还必须研究在系统中设计正义和公平因素的程度，以及如何在当前工程技术的范围内实现这一点。

建立道德AI

除了正义和公平的基本假设之外，人工智能系统是否能够表现出遵守一般道德原则的行为。人工智能的进步如何在道德规范中提出新的"机器相关"问题，或者人工智能的使用可能被认为是不道德的？道德本质上是一个哲学问题，而人工智能技术依赖于工程学，并受其限制。因此，在技术可行的范围内，研究人员必须努力开发可验证地符合或符合现有法律，社会规范和道德规范的算法和架构 - 显然是一项非常具有挑战性的任务。道德原则通常具有不同程度的模糊性，难以转化为精确的系统和算法设计。当AI系统，尤其是新型自主决策算法，面临基于独立且可能相互冲突的价值体系的道德困境时，也会出现复杂情况。道德问题因文化，宗教和信仰而异。但是，可以制定可接受的道德参考框架来指导人工智能系统的推理和决策，以便解释和证明其结论和行动的合理性。需要采用多学科方法来生成反映适当价值体系的培训数据集，包括表明在遇到困难的道德问题或具有相互冲突的价值时的首选行为的例子。这些示例可以包括合法或道德的"极端案例"，由对用户透明的结果或判断标记。[68] 人工智能需要适当的方法来解决基于价值的冲突，其中系统包含的原则可以解决严格规则不切实际的复杂情况的现实。

为道德AI设计架构

必须在基础研究方面取得进一步进展，以确定如何最好地设计包含道德推理的AI系统架构。已经提出了各种方法，例如将操作AI与监视器代理分开的双层监视器体系结构，监视器代理负责对任何操作操作进行道德或法律评估。.[68] 另一种观点是安全工程是首选，其中使用AI代理体系结构的精确概念框架来确保AI行为是安全的并且对人类无害。[69] 第三种方法是使用集合理论原则与逻辑约束相结合来制定道德体系结构

[68] A. Etzioni和O. Etzioni，"设计符合我们法律和价值观的人工智能系统"，ACM通讯59（9）（2016）：29-31。

[69] RY Yampolsky，"人工智能安全工程：为什么机器伦理是一种错误的方法。"在哲学和人工智能理论，编辑。VC Muller（Heidelberg：Springer Verlag，2013），389-396。

关于限制行动以符合道德原则的AI系统行为。[70] 随着人工智能系统变得更加通用，他们的架构可能会包含可以在多个判断级别上处理道德问题的子系统，包括：[71] 快速反应模式匹配规则，用于描述和证明行动的较慢响应的协商推理，用于指示用户可信度的社交信号，以及在更长时间尺度上操作以使系统遵守文化规范的社交过程。研究人员需要关注如何最好地解决符合道德，法律和社会目标的AI系统的整体设计。

[70]RC Arkin，“管理法律行为：在混合协商/反应机器人架构中嵌入道德”，佐治亚理工学院技术报告，GIT-GVU-07-11,2007。

[71]B.Kuipers，“人类道德与机器人伦理”，AAAI-16人工智能，伦理与社会研讨会，2016年;https://web.eecs.umich.edu/~kuipers/papers/Kuipers-aaaiws-16.pdf

战略4：确保AI系统的安全性

2019更新	创建健壮且值得信赖的AI系统

自2016年国家AI研发战略计划发布以来，人工智能安全和安全的科学和社会理解迅速增长。这些新知识中的大部分都有助于发现新问题：现在更明显的是，人工智能系统如何能够做错事，学习错误信息或揭露错误信息，例如，通过对抗性示例，数据中毒和模型反演分别。不幸的是，这些人工智能安全和安全问题的技术解决方案仍然难以捉摸。

为解决所有这些问题，必须在AI系统生命周期的所有阶段（从初始设计和数据/模型构建，到验证和验证，部署，操作和监控）考虑AI系统的安全性和安全性。实际上，“设计安全（或安全）”的概念可能会产生一种错误的观念，即这些只是系统设计者的关注点;相反，它们必须在整个系统生命周期中考虑，而不仅仅是在设计阶段，因此必须是AI研发组合的重要组成部分。

当AI组件连接到其他系统或必须安全或安全的信息时，AI漏洞和性能要求（例如，当在大量数据上操作时，误报率和假阴性率非常低）

[72] https://www.transportation.gov/av/3
[73] https://www.nhtsa.gov/automated-driving-systems/v
[74] https://www.iarpa.gov/index.php/research-programs/
[75] https://www.iarpa.gov/index.php/research-programs/
[76] https://www.darpa.mil/news-events/2019-02-06
[77] “2018年国防部人工智能战略总结”：https://media.d
1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.

人工智能安全与保障：最近的代理机构研发计划

自2016年国家AI研发战略计划发布以来，一些机构已开始支持战略4：

- DOT于2018年10月发布了新的联邦自动驾驶汽车指南，支持将自动化安全地集成到广泛的多式联运系统中。为交通的未来做好准备：自动驾驶汽车3.0[72] 推进DOT的自动驾驶汽车安全整合原则。该文件还重申了先前的安全指南，提供了新的多模式安全指导，并概述了随着这项新技术的发展而与DOT合作的过程。截至2019年5月，已有14家公司发布了自愿安全自我评估，详细说明了如何将安全性纳入其自动驾驶系统的设计和测试中。[73]
- 在2018年12月，IARPA宣布了两项关于人工智能安全的计划：安全，可靠，智能学习系统（SAILS）[74] 和人工智能中的特洛伊木马（TrojAI）。[75] DARPA于2019年2月宣布另一项计划，即保证人工智能对欺骗行为的可靠性（GARD）。[76] 这些计划旨在共同打击对AI系统的一系列攻击。
- 如战略3所述，国防部致力于“引领军事道德和人工智能安全”，作为指导其加速采用人工智能系统的战略方法中列出的五项关键行动之一。[77]

由较大的系统继承。这些挑战并非一成不变;随着人工智能系统能力的不断提高，他们的复杂性可能会越来越高，这使得更难以正确的性能或信息的隐私得到验证和验证。这种复杂性也可能使得越来越难以以证明人类用户高度信任的方式解释决策（见战略3）。

让AI值得信赖 – 现在和未来 – 是一个关键问题，需要联邦政府的研发投资（见边栏），以及政府，行业，学术界和民间社会之间的合作。工程值得信赖的AI系统可以从其他领域的安全工程中借用现有实践中受益，这些领域已经学会了如何解释非AI自主或半自治系统的潜在不当行为。然而，特定于AI的问题意味着用于程序分析，测试，形式验证和综合的新技术对于确定基于AI的系统满足其规范（即系统完全按照预期执行的操作）至关重要。而且没有了。这些问题在基于人工智能的系统中会加剧，这些系统很容易被愚弄，逃避和误导，其方式可能具有深远的安全隐患。一个新兴的研究领域是对抗性ML，它探讨了ML算法中的漏洞分析以及产生更强大学习的算法技术。对ML的众所周知的攻击包括对抗性分类器规避攻击，其中攻击者改变行为以逃避被检测，以及中毒攻击，其中训练数据本身被破坏。越来越需要研究系统地探索攻击ML和其他基于AI的系统的攻击者的空间，并设计能够提供针对攻击类别的可证实的鲁棒性保证的算法。

必须开发方法以确保AI的创建，评估，部署和控制的安全性，并且这些方法必须扩展以匹配AI的能力和复杂性。评估这些方法将需要新的指标，控制框架和基准，以测试和评估日益强大的系统的安全性。方法和指标都必须包含人为因素，人类设计者目标定义的安全AI目标，人类用户习惯定义的安全AI操作，以及人类评估者理解的安全AI指标。为人工智能系统的安全性制定人为驱动和人类可理解的方法和指标将使政策制定者，私营部门和公众能够准确地判断不断变化的人工智能安全格局并在其中适当地进行。

在人工智能系统得到广泛使用之前，需要确保系统以受控方式安全可靠地运行。需要进行研究以解决创建可靠，可靠和值得信赖的AI系统的挑战。与其他复杂系统一样，AI系统面临重要的安全和安全挑战，原因如下：[78]

- *复*杂和不确定的环境：在许多情况下，AI系统被设计为在复杂环境中运行，具有大量潜在状态，无法对其进行详尽检查或测试。系统可能面临在设计过程中从未考虑过的条件。
- *紧*急行为：对于在部署后学习的AI系统，系统的行为可能主要由在无监督条件下学习的时段决定。在这种情况下，可能难以预测系统的行为。
- *目*标错误指定：由于难以将人类目标转换为计算机指令，因此为AI系统编程的目标可能与程序员所预期的目标不匹配。

[78] J. Bornstein, "DoD Autonomy Roadmap – Autonomy Community of Interest", NDIA第16届年度科学与工程技术会议上的演讲，2015年3月。

- 人机交互：在许多情况下，人工智能系统的性能受人类交互的影响很大。在这些情况下，人类反应的变化可能会影响系统的安全性。[79]

为了解决这些问题和其他问题，需要额外的投资来推进人工智能的安全和保障，[80] 包括可解释性和透明度，信任，验证和验证，针对攻击的安全性以及长期AI安全性和价值调整。

提高可解释性和透明度

一项关键的研究挑战是提高AI的“可解释性”或“透明度”。许多算法（包括基于深度学习的算法）对用户来说是不透明的，几乎没有用于解释其结果的现有机制。对于诸如医疗保健之类的领域而言，这尤其成问题，其中医生需要解释来证明特定诊断或治疗过程的合理性。诸如决策树感应之类的AI技术提供内置解释，但通常不太准确。因此，研究人员必须开发透明的系统，并且本质上能够向用户解释其结果的原因。

建立信任

为了获得信任，AI系统设计人员需要创建具有信息性，用户友好界面的准确，可靠的系统，而操作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，例如车辆的手动控制，往往是透明的（系统以用户可见的方式运行），可信（系统的输出被用户接受），可审计（系统可以被评估），可靠（系统充当用户的意图）和可恢复（用户可以在需要时恢复控制）。当前和未来AI系统面临的重大挑战仍然是软件生产技术的质量不稳定。随着进步带来人与人工智能系统之间更大的联系，信任领域的挑战是跟上变化和增加能力的步伐，预测采用和长期使用的技术进步，并为最佳研究制定治理原则和政策。设计，施工和使用的实践，包括安全操作的适当操作员培训。

加强验证和验证

AI系统的验证和验证需要新方法。“验证”确定系统符合正式规范，而“验证”确定系统满足用户的操作需求。安全的AI系统可能需要新的评估方法（确定系统是否出现故障，可能是在超出预期参数的情况下运行），诊断（确定故障原因）和维修（调整系统以解决故障）。对于在长时间内自主运行的系统，系统设计人员可能没有考虑系统将遇到的每种情况。此类系统可能需要具备自我评估，自我诊断和自我修复的能力，以便稳健可靠。

---

[79] JM Bradshaw，RR Hoffman，M。Johnson和DD Woods，“自治系统的七大致命神话”，*IEEE Intelligent Systems 28 (3) (2013) : 54-61*。

[80] 例如：见D. Amodei，C。Olah，J。Steinhardt，P。Christiano，J。Schulman和D. Mane，“人工智能安全中的具体问题”，2016年，ARXIV: 160606565V2; S. Russell，D。Dewey和M. Tegmark，“稳健和有益的人工智能的研究重点”，2016年，arXiv:1602.03506;TG Dietterich和EJ Horvitz，“对AI的关注的兴起：反思与方向”，ACM通讯，58（10）（2015）;和K. Sotala和R. Yampolskiy，“应对灾难性AGI风险：一项调查”，Physica Scripta，90（1），2014年12月19日。

防范攻击

嵌入关键系统的AI必须是强大的，以便处理事故，但也应该对各种有意的网络攻击安全。安全工程涉及了解系统的漏洞以及可能有兴趣攻击它的行为者的行为。在NITRD 2016联邦网络安全研发战略计划中更详细地讨论了网络安全研发需求，[81] 一些网络安全风险是AI系统特有的。例如，一个关键的研究领域是“对抗性机器学习”，它探讨了通过“污染”训练数据，修改算法或对对象进行细微更改以防止其被正确识别而影响AI系统的程度。 （例如，欺骗面部识别系统的假肢）。在需要高度自治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。最近在这一领域工作的一个例子是DARPA的网络大挑战，其中涉及AI代理人自主分析和打击网络攻击。[82]

实现长期人工智能安全和价值调整

AI系统最终可能具有“递归自我改进”的能力，其中软件本身而不是人类程序员进行了大量的软件修改。为了确保自我修改系统的安全性，需要进一步开展研究：自我监控架构，检查系统与人类设计师的原始目标的行为一致性；禁止在评估系统时释放系统的限制策略；价值学习，其中用户的价值观，目标或意图可以由系统推断；和价值框架，可以证明自我修改。

[81] https://www.nitrd.gov/pubs/2016-Federal-Cybersecurity-Research-and-Development-Strategic-Plan.pdf; 这是在2019年更新。

[82] https://archive.darpa.mil/CyberGrandChallenge_CompetitorSite/

策略5：为人工智能培训和测试开发共享的公共数据集和环境

2019 更新	增加对数据集和相关挑战的访问

在2016年国家AI研发战略计划发布时，公开可用的数据集和环境已经在推动AI研发方面发挥了关键作用，特别是在计算机视觉，自然语言处理和语音识别等领域。ImageNet[84] 超过1400万个标记对象，以及相关的计算机视觉社区挑战（例如，ImageNet大规模视觉识别挑战[85] 评估对象检测和图像分类的算法），在社区中发挥了特别重要的作用。随着ML的转化应用程序在医疗保健，医学，智能和互联社区等众多应用领域中被发现，对特定领域领域的公开数据集的需求也在增长。

数据集和模型的重要性 – 特别是联邦政府的数据集和模型 – 在2019年维持美国人工智能领导力行政命令中明确提出:[1]

> 所有机构的负责人都应审查其联邦数据和模型，以确定更大的非联邦AI研究社区以有利于该社区的方式增加访问和使用的机会，同时保护安全，安全，隐私和机密性。具体而言，代理商应改进数据和模型库存文档以实现发现和可用性，并且应该

[83] https://insight.shrp2nds.us/
[84] http://www.image-net.org/
[85] http://www.image-net.org/challenges/LSVRC/

用于AI培训和测试的共享公共数据集和环境：最近的代理研发计划

自2016年国家人工智能研发战略计划发布以来，一些机构已开始支持战略5：

- DOT赞助了第二次战略公路研究计划（SHRP2）自然驾驶研究（NDS），[83]记录超过超过3,400名司机和车辆乘坐540万人次。车载数据采集系统（DAS）单元收集并存储来自前方雷达，四个摄像机，加速度计，车辆网络信息，地理定位系统和车载车道跟踪器的数据。在参与者的车辆运行时，DAS的数据不断记录。虽然NDS数据的摘要是公开的，但访问详细数据集需要合格的研究伦理培训。
- VA Data Commons正在创建世界上最大的链接医学基因组数据集，其中包含用于启用ML和AI的工具，并由退伍军人的偏好指导。这项工作正在利用适用的NIST标准，法律和行政命令。
- GSA（一般服务管理局）正致力于将云计算资源用于联邦政府资助的AI研发。位于GSA的Data.gov和code.gov包含来自各机构的超过246,000个数据集和代码，并自动收集各机构发布的数据集。
- NIH发现，实验和可持续发展的科学和技术研究基础设施（STRIDES）倡议与行业领先的云服务提供商建立了合作关系，使研究人员能够访问由NIH资助并存储在云环境中的主要数据资产。

根据AI研究社区的反馈，优先考虑改进AI数据和模型的访问和质量。

新的NSTC开放科学小组委员会于2018年成立，旨在协调联邦在开放和FAIR（可查找，可访问，可互操作和可重用）数据方面的努力。将需要研发投资来开发工具和资源，以便更容易识别，使用和操纵相关数据集（包括联邦数据集），验证数据来源，并尊重适当的使用政策。许多这些数据集本身在人工智能环境中的使用可能有限，而无需在标签和管理方面进行投资。联邦机构应该与AI利益相关方合作并确保适当审查的数据集和模型已经准备好并且适合使用，并且随着标准和规范的发展而得到维护。最终，在记录数据集和模型出处时开发和采用最佳实践和标准将提高可靠性和负责任地使用AI技术。

自2016年以来，人们对数据内容的担忧也越来越多，例如潜在的偏见（见战略3）[86,87] 或私人信息泄露。2016年国家AI研发战略计划指出，“数据集的开发和共享必须遵循适用的法律法规，并以道德的方式进行。”DOT支持的InSight项目提供了对自然驾驶研究期间收集的数据的精心组织的访问。 （见附文）。2016年国家AI研发战略计划还指出，需要新的“技术来确保数据的安全共享，因为数据所有者在与研究团体共享数据时承担风险。”例如，CryptoNets[88] 允许神经网络对加密数据进行操作，确保数据保密，因为神经网络中不需要解密密钥。研究人员还开始开发新的ML技术，使用差异隐私框架为使用的数据提供可量化的隐私保证。[89] 同时，隐私方法必须保持足够的可解释性和透明度，以帮助研究人员纠正它们并使其安全，有效和准确。此外，AI可以揭示超出原始或预期范围的发现;因此，研究人员必须认识到对手获取数据或发现的潜在危险。

如果没有将计算资源用于大规模公共数据集的能力，那么单独使用数据几乎没有用处。计算资源对人工智能研发的重要性在2019年“维持美国人工智能领导力行政命令”中得到了体现:[1]

> 国防，商业，卫生和人类服务部门和能源部长，美国国家航空航天局局长和国家科学基金会主任应在适当和适用的范围内，优先考虑高分配 - 通过以下方式为AI相关应用程序提供的性能计算资源： （i）增加对资源和资源储备的自由分配的分配;或（ii）任何其他适当的机制。

[86]Emily M. Bender和Batya Friedman，“NLP的数据陈述：趋向于减轻系统偏差和实现更好的科学”，计算语言学协会的交易6（2018）：587-604。

[87]Aylin Caliskan，Joanna J. Bryson和Arvind Narayanan，“自动从语料库中衍生的语义包含类似人类的偏见”，Science 356（6334）：183-186，2017年4月14日。

[88]Ran Gilad-Bachrach，Nathan Dowlin，Kim Laine，Kristin Lauter，Michael Naehrig，John Wernsing，“CryptoNets：将神经网络应用于高吞吐量和高精度的加密数据”，2016年国际机器学习会议48：201-210 ；http://proceedings.mlr.press/v48/.

[89]Martin Abadi，Andy Chu，Ian Goodfellow，H。Brendan McMahan，Ilya Mironov，Kunal Talwar和Li Zhang，“深度学习与差异隐私”，第23届ACM计算机与通信安全会议，“2016：308-318 。

和：

> .....专责委员会应与总务管理局（GSA）协调，向总统提交报告，就更好地使用云计算资源进行联邦政府资助的人工智能研发提出建议。

对许多AI挑战的计算能力的需求一直在迅速增加.[32] 联邦资助可为联邦政府资助的研究提供计算能力。然而，一些公司和大学可能有额外的计算需求。总体而言，国家需要研究和投资共享计算资源以促进人工智能研发。

人工智能的好处将继续增加，但仅限于开发和提供人工智能的培训和测试资源。训练数据集和其他资源的多样性，深度，质量和准确性会显着影响AI性能。许多不同的AI技术需要高质量的数据用于培训和测试，以及动态，交互式测试平台和模拟环境。不仅仅是一个技术问题，这是一个重要的"公益"挑战，因为如果人工智能培训和测试仅限于少数已经拥有宝贵数据集和资源的实体，那么进步将受到影响，但我们必须同时尊重商业和个人权利和对数据的兴趣。需要开展研究，为各种AI应用开发高质量的数据集和环境，并负责访问良好的数据集，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研发的进程。以下小节概述了这些重要的关键领域。

开发和提供各种数据集，以满足各种AI兴趣和应用的需求

人工智能培训和测试数据集的完整性和可用性对于确保科学可靠的结果至关重要。支持数字领域可重复研究所需的技术和社会技术基础设施已被认为是一项重要挑战 – 对人工智能技术也至关重要。缺乏经过审查和公开可用的数据集以及确定的可重复性的数据集是人工智能自信进步的关键因素。[90] 与其他数据密集型科学一样，捕获数据来源至关重要。研究人员必须能够使用相同和不同的数据集重现结果。数据集必须代表具有挑战性的真实世界应用程序，而不仅仅是简化版本。为了快速取得进展，应该把重点放在提供政府持有的现有数据集，可以用联邦资金开发的数据集，以及尽可能用工业持有的数据集。

AI挑战的机器学习方面通常与"大数据"分析相关联。考虑到各种相关数据集，对非结构化或半结构化数据进行适当的表示，访问和分析仍然是一个日益严峻的挑战。如何以绝对和相对（依赖于上下文）的术语表示数据？当前的真实数据库可能非常容易受到不一致，不完整和嘈杂的数据的影响。因此，许多数据预处理技术（例如，数据清理，集成，转换，缩减和表示）对于为AI应用程序建立有用的数据集非常重要。数据预处理如何影响数据质量，尤其是在执行其他分析时？

[90]为此，2016年，情报高级研究项目活动发布了关于新型训练数据集和环境的信息请求，以推进AI。看到 https://iarpa.gov/index.php/working-with-iarpa/请求换信息/新颖，训练数据集和的环境中对提前人工智能。

鼓励共享AI数据集 – 尤其是政府资助的研究 – 可能会激发创新的AI方法和解决方案。但是，需要技术来确保数据的安全共享，因为数据所有者在与研究社区共享数据时会承担风险。数据集开发和共享还必须遵守适用的法律和法规，并以道德的方式进行。风险可能以各种方式出现：不恰当地使用数据集，不准确或不恰当的披露，以及数据去识别技术的限制，以确保隐私和机密性保护。

根据商业和公共利益制定培训和测试资源

随着全球数据，数据源和信息技术的不断爆炸，数据集的数量和规模都在不断增加。分析数据的技术和技术无法跟上大量原始信息源。数据采集，策展，分析和可视化都是关键的研究挑战，从大量数据中提取有价值的知识所需的科学是滞后的。虽然存在数据存储库，但它们通常无法处理数据集的扩展，数据来源信息有限，并且不支持语义丰富的数据搜索。需要动态，灵活的存储库。

支持人工智能研究需求所需的开放/共享基础设施计划的一个例子是国土安全部（DHS）制定的IMPACT计划（网络风险与信任政策与分析信息市场）。[91] 该计划通过协调和开发实际数据和信息共享功能（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经验数据共享。AI研发将受益于所有AI应用程序中的可比程序。

开发开源软件库和工具包

开源软件库和工具包的可用性越来越高，可以为任何具有Internet连接的开发人员提供基于图像的AI技术。Weka工具包等资源，[92] 槌，[93] 和OpenNLP，[94] 在许多其他人中，加速了人工智能的开发和应用。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）以及免费或低成本的开发语言（例如，R，Octave和Python）为使用和扩展这些库提供了较低的障碍。此外，对于那些可能不想直接集成这些库的人，可以使用任意数量的基于云的机器学习服务，这些服务可以通过需要很少或不需要编程的低延迟Web协议按需执行图像分类。最后，许多这些Web服务还提供专用硬件的使用，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）也将通过这些服务广泛使用。

这些资源共同提供了一个AI技术基础设施，鼓励市场创新，允许企业家开发解决方案，解决窄域问题，无需昂贵的硬件或软件，无需高水平的AI专业知识，并允许按需快速扩大系统规模。对于狭窄的AI领域，与许多其他技术领域相比，市场创新的障碍极低。

---

[91] https://www.dhs.gov/csd-impact

[92] https://sourceforge.net/projects/weka/

[93] http://mallet.cs.umass.edu

[94] https://opennlp.apache.org

为了帮助支持该领域持续的高水平创新，美国政府可以加大开放，支持和使用开放式人工智能技术的力度。特别有益的是使用标准化或开放格式的开放资源和用于表示语义信息的开放标准，包括可用时的领域本体。

政府还可以通过加速在政府内部使用开放式人工智能技术来鼓励更多地采用开放的人工智能资源，从而有助于保持创新者进入门槛较低。政府应尽可能为开源项目提供算法和软件。由于政府有特定的关注点，例如更加重视数据隐私和安全性，政府可能有必要制定机制来减轻政府采用人工智能系统的速度。例如，创建一个可以跨政府机构执行“水平扫描”以找到部门内特定AI应用领域的工作组，然后确定需要解决的特定问题以允许采用此类技术可能是有用的。这些机构。

策略6：通过标准和基准测量和评估AI技术

2019更新 支持开发人工智能技术标准和相关工具

2016年国家人工智能研发战略计划指出，“人工智能社区的标准，基准，测试平台及其采用对于指导和促进人工智能技术的研发至关重要。”在此期间的三年中，对标准和基准的重视程度不断提高在美国和全球。2019年“维持美国人工智能领导力”行政命令明确提出了这些标准的重要性:[1]

> .....商务部长通过[NIST]主任，应制定联邦参与技术标准和相关工具开发的计划，以支持使用AI技术的可靠，稳健和可靠的系统。

随着人工智能创新可能影响社会的所有部门和领域，许多标准制定组织正在进行新的与人工智能相关的考虑和工作项目，包括与人工智能道德和可信赖的人工智能系统相关的活动（见战略3）。国际标准化组织（ISO）和国际电工委员会（IEC）召集了一个关于AI的联合技术小组委员会（ISO / IEC联合技术委员会1，小组委员会）42关于人工智能[95]）制定人工智能系统标和学术研究人员应继续为这些活动提供信机构的使命相一致的交通，医疗保健和食

自2016年以来，人工智能相关标准活动的的推出，特别是与人工智能系统的可信度

标准，基准和相关工具：最近的机构研发计划

自2016年国家AI研发战略计划发布以来，NIST特别为战略6开展了工作：

- NIST参与ISO / IEC JTC 1 SC 42关于人工智能的标准化计划。[95] NIST专家是SC 42中大数据工作的召集人。美国驻SC 42代表团包括NIST和其他联邦机构专家，以及来自工业界和学术界的代表。美国对SC 42的投入由国际信息技术标准委员会（INCITS）提供便利。
- NIST员工通过标准组织参与其他AI标准活动，例如美国机械工程师协会，IEEE和ISO / IEC。他们的活动涉及诸如先进制造的计算建模，机器人和自动化的本体，个人数据隐私和算法偏差等主题。
- NIST专家正在提高人们对多边论坛中人工智能共识标准重要性的认识，包括G20和G7等机构。[96] NIST带来了独特的联邦政府专业知识，可以在实践中进行政策讨论，特别是通过与私营部门的密切合作。同样，NIST将其标准和相关经验用于政府间双边讨论。

[95] https://www.iso.org/committee/6794475.html
[96] https://home.treasury.gov/policy-issues/international/g-7-and-g-20

然而，干预时间，基准数据集中的公平性和偏见的考虑变得越来越重要，研究人员正在寻求新的面部识别数据集，以寻求最小化偏差。测试AI算法的应用程序级性能（例如，分类算法的假阳性或假阴性率）和量化AI软件和硬件系统的计算级性能的基准测试的基准测试更为丰富。最近的两项活动是MLPerf[97]和DAWNbench。[98]

评估，促进和确保AI可信赖性的所有方面需要通过基准和标准来衡量和评估AI技术性能。除了安全，可靠，可靠，有弹性，可解释和透明之外，值得信赖的AI必须保护隐私，同时检测并避免不适当的偏见。随着人工智能技术的发展，开发新指标和测试要求以验证这些基本特征的需求也将随之增加。

AI社区的标准，基准，测试平台及其采用对于指导和促进AI技术的研发至关重要。以下小节概述了必须取得进一步进展的领域。

制定广泛的AI标准

必须加快标准的开发，以跟上快速发展的能力和扩展AI应用领域的步伐。标准提供了可以一致使用的要求，规范，指南或特性，以确保AI技术满足功能和互操作性的关键目标，并且可靠且安全地执行。采用标准为技术进步带来可信度，并促进扩展的可互操作市场。已经开发的AI相关标准的一个示例是由电气和电子工程师协会开发的P1872-2015（机器人和自动化的标准本体）。该标准提供了表示知识的系统方法和一组通用的术语和定义。这些允许在人类，机器人和其他人工系统之间进行明确的知识转移，并为人工智能技术应用AI技术提供基础。人工智能的所有子域都需要在AI标准开发方面进行额外的工作。

需要标准来解决：

- *软*件工程：管理系统复杂性，维护，安全性以及监控和控制紧急行为；
- *性*能：确保准确性，可靠性，稳健性，可访问性和可扩展性；
- *度*量标准：量化影响性能和符合标准的因素；
- *安*全性：评估系统的风险管理和危害分析，人机交互，控制系统和法规遵从性；
- *可*用性：确保界面和控件的有效性，高效性和直观性；
- *互*操作性：通过标准和兼容接口定义可互换的组件，数据和事务模型；
- *安*全性：解决信息的机密性，完整性和可用性以及网络安全问题；
- *隐*私：控制在处理，运输或存储时保护信息；

[97] https://mlperf.org/

[98] https://dawn.cs.stanford.edu/benchmark/

- *可*追溯性：提供事件记录（实施，测试和完成）以及数据管理;和
- *域*：定义特定于域的标准词典和相应的框架。

建立AI技术基准

由测试和评估组成的基准为制定标准和评估标准的合规性提供了量化指标。基准通过促进旨在解决战略性选择方案的进步来推动创新;他们还提供客观数据来跟踪人工智能科学和技术的发展。为了有效地评估AI技术，必须开发和标准化相关且有效的测试方法和指标。标准测试方法将规定评估，比较和管理AI技术性能的协议和程序。需要使用标准指标来定义可量化的度量，以便表征AI技术，包括但不限于：准确性，复杂性，信任和能力，风险和不确定性，可解释性，意外偏差，与人类绩效的比较以及经济影响。值得注意的是，基准测试是数据驱动的。策略5讨论了数据集对培训和测试的重要性。

作为与AI相关的基准测试的成功范例，美国国家标准与技术研究院开发了一套全面的标准测试方法和相关的性能指标，以评估应急响应机器人的关键功能。目的是通过利用使用标准测试方法捕获的机器人能力的统计上重要的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功能。通过ASTM国际安全应用标准委员会对机器人操作设备（称为标准E54.08.01）进行了标准化测试。[99] 测试方法的版本用于通过RoboCup救援机器人联盟比赛挑战研究社区，[100] 强调自主能力。另一个例子是IEEE Agile Robotics for Industrial Automation Competition（ARIAC），[101] IEEE与NIST的共同努力，[102] 通过利用人工智能和机器人规划的最新进展，提高机器人的灵活性。本次竞赛的核心焦点是测试工业机器人系统的灵活性，目标是使车间的人员更高效，更自主，并且需要更少的车间工人时间。

虽然这些努力为推动人工智能基准测试提供了坚实的基础，但它们受到特定领域的限制。更广泛的领域需要额外的标准，测试平台和基准，以确保AI解决方案广泛适用并广泛采用。

增加AI测试平台的可用性

“未来的网络实验”报告中指出了测试平台的重要性：“测试平台是必不可少的，以便研究人员可以使用实际的操作数据来模拟和运行真实系统的实验.....以及良好测试环境中的情景。”[103] 有足够的测试台是一个

---

[99] 2019更新：由此产生的测试方法现在是ASTM国际标准委员会在国土安全响应机器人应用程序中发布的标准（称为E54.09）。

[100] http://www.robocup2016.org/en/

[101] http://robotagility.wixsite.com/competition

[102] 2019年更新：IEEE不再是ARIAC的合作伙伴，现已进入第三年。

[103] SRI国际和南加州大学信息科学研究所，“未来网络安全实验（CEF）：催化新一代实验性网络安全研究”，最终报告，2015年7月31日。

需要跨越AI的所有领域。政府拥有大量政府独有的任务敏感数据，但大部分数据无法分发给外部研究界。可以为学术和工业研究人员建立适当的计划，以便在特定机构建立的安全和策划的测试平台环境中进行研究。AI模型和实验方法可以由研究团体通过访问这些测试环境来共享和验证，为AI科学家，工程师和学生提供独特的研究机会。

让AI社区参与标准和基准测试

需要政府的领导和协调来推动标准化并鼓励其在政府，学术界和工业界广泛使用。由用户，行业，学术界和政府组成的AI社区必须充满活力，参与制定标准和基准计划。由于每个政府机构根据其角色和使命以不同方式参与社区，因此可以通过协调来利用社区互动，以加强其影响。需要这种协调来集体收集用户驱动的需求，预测开发人员驱动的标准，并促进教育机会。用户驱动的需求决定了挑战问题的目标和设计，并使技术评估成为可能。拥有社区基准专注于研发，以确定进度，缩小差距，并针对特定问题推动创新解决方案。这些基准必须包括定义和分配基本事实的方法。基准模拟和分析工具的创建也将加速AI的发展。这些基准测试的结果还有助于将正确的技术与用户的需求相匹配，形成符合标准的客观标准，合格的产品列表和潜在的源选择。

工业界和学术界是新兴AI技术的主要来源。促进和协调他们参与标准和基准活动至关重要。随着解决方案的出现，通过分享技术架构的共同愿景，开发新兴标准的参考实现以显示可行性，以及进行预竞争性测试以确保高质量和可互操作的解决方案，以及为了预测开发人员和用户驱动的标准，有很多机会。开发技术应用的最佳实践。

高度影响，基于社区，与AI相关的基准程序的一个成功范例是文本检索会议（TREC），[104] 这是由NIST于1992年启动的，旨在提供大规模评估信息检索方法所需的基础设施。超过250个团体参加了TREC，包括大小学术和商业组织。TREC提出的标准的，广泛可用的，精心构建的数据集已被用于振兴信息检索研究。[105,106] 第二个例子是应用于生物识别技术的机器视觉领域的NIST定期基准程序，[107] 特别是面部识别。[108] 这开始于1993年的人脸识别技术（FERET）评估，该评估提供了面部照片的标准数据集，旨在支持人脸识别算法开发以及评估协议。多年来，这项努力已经发展成为Face

[104] http://trec.nist.gov

[105] EM Voorhees和DK Harman，TRECExstrationment and Evaluation in Information Retrieval（Cambridge：MIT Press，2005）。

[106] http://googleblog.blogspot.com/2008/03/why-data-matters.html

[107] http://biometrics.nist.gov

[108] http://face.nist.gov

认可供应商测试（FRVT），[109] 涉及数据集的分发，托管挑战问题以及进行隔离技术评估。该基准程序为面部识别技术的改进做出了巨大贡献。TREC和FRVT都可以作为有效的AI相关社区基准活动的例子，但AI的其他领域也需要类似的努力。

值得注意的是，制定和采用标准以及参与基准活动需要付出代价。研发组织在看到重大利益时会受到激励。更新各机构的采购流程，以便在提案请求中包含AI标准的特定要求，这将鼓励社区进一步参与标准制定和采用。基于社区的基准，如TREC和FRVT，还通过提供各种类型的培训和测试数据来降低障碍并加强激励，否则无法获取，促进技术开发人员之间的良性竞争，以推动最佳算法，并提供客观和比较性能指标用于相关的源选择。

[109]PJ Phillips，“改善人脸识别技术”，计算机44（3）（2011）：84-96。

战略7：更好地了解国家AI研发人员的需求

2019 更新	推动人工智能研发人员，包括那些致力于人工智能系统和与他们一起工作的人员，以维持美国的领导地位

自2016年国家AI研发战略计划发布以来，对AI研究人员和从业人员的需求迅速增长。研究表明，在未来十年内，招聘机会的数量有望增加到数百万。作为一个数据点，美国劳工统计局预计，从2016年到2026年，计算机和信息科学家和工程师的职位数量将增长19%，几乎是所有职业平均数的三倍。[111] 此外，到2028年，人工智能研究人员预计仅贡献于G20国家智能技术承诺的11.5万亿美元的累积增长。[112]

美国学术机构正在努力跟上学生兴趣和人工智能入学的爆炸性增长。[113,114,115] 与此同时，行业凭借其持续的财务支持以及对先进计算设施和数据集的访问，对学术研究和教学人才产生了强大的影响。[116]

在人工智能中保持强大的学术研究生态系统至关重要，与生产者研发部门合作，可以继续带来巨大的回报[117] 通过促进国民健康，繁荣和福利，确保国防。

国家人工智能研发人员：近期的机构活动

自2016年国家AI研发战略计划发布以来，一些机构已开始支持战略7：

- 除了通过标准的人工智能研究资助支持本科生和研究生之外，各机构还在其研究生奖学金计划中优先考虑计算和数据支持的科学和工程。例如，在2018年，DOE为其计算科学研究生奖学金计划增加了一条新的轨道。该课程支持学生攻读应用数学，统计学或计算机科学的高级学位，并促进更有效地使用高性能系统，包括AI，ML和深度学习领域.[44,110] 同样在2018年，NSF开始在其研究生研究奖学金计划的获奖者的子集中优先考虑计算和数据支持的科学和工程。
- 人口普查局创建了统计数据现代化（SDM）项目，以使其劳动力，运营和技术达到当前的最新水平，并为当今数据驱动型社会中的统计机构设定标准。SDM的劳动力转换组件将雇用具有新方法和分析专业知识的新数据科学家，包括使用AI方法和工具来处理和分析大数据。劳动力转型还将解决当前数据科学人员的技能提升问题。

[110] https://www.krellinst.org/csgf/math-cs
[111] https://www.bls.gov/ooh/computer-and-information
[112] HTTPS://www. ACCENTURU.COM/T20180920T094705Z /埃森哲教育和技术技能，研究.pdf
[113] https://cra.org/data/generation-cs/
[114] https://cra.org/wp-content/uploads/2018/05/2017-T
[115] http://web.cs.wpi.edu/~cew/papers/CSareas19.pdf
[116] https://www.nitrd.gov/rfi/ai/2018/AI-RFI-Response-2
[117] https://www.nap.edu/catalog/13427/continuing-inno

自2016年国家人工智能研发战略计划发布以来的三年中，各种报告呼吁继续支持各级计算机科学教材和教师专业发展的发展。几十年来，K-12层面需要强调为国家人工智能研究人员提供支持。[118] 在本科阶段，考虑到跨学科的计算越来越重要，需要将重点放在将先进的计算技能和方法与来自其他学科的特定领域知识相结合。[119] 研究生阶段也需要持续的支持，学生们正在进行ML和AI的基础研究。实际上，2019年维持美国人工智能领导力的行政命令要求这样做:[1]

> 提供教育补助金的执行机构负责人应在符合适用法律的范围内，将人工智能视为现有联邦奖学金和服务计划中的优先领域.....［包括］.....（A）高中，本科和研究生奖学金;替代教育;和培训计划；（B）承认和资助进行人工智能研发的早期职业大学教师的计划，包括通过总统奖励和表彰；（C）服务计划奖学金；（D）美国武装部队的直接调试方案；（E）支持制定教学计划和课程的计划，鼓励将人工智能技术纳入课程，以促进正规和非正规教育和培训的个性化和适应性学习体验。

更广泛地说，在联邦政府2018年12月的科学，技术，工程和数学（STEM）教育五年战略计划中，也突出了对计算思维的坚实基础的需求，包括通过计算机科学教育。[120]

此外，必须扩大传统上在计算机和相关领域中代表性不足的群体的参与。

最后，AI研发人员将由多学科团队组成，不仅包括计算机和信息科学家和工程师，还包括其他领域的专家，这些领域是人工智能和ML创新及其应用的关键，包括认知科学和心理学，经济学和博弈论，工程学和控制理论，伦理学，语言学，数学，哲学以及可能应用AI的许多领域。

联邦机构优先考虑各级培训和奖学金计划，通过学徒，技能计划，奖学金和相关学科的课程工作，为员工提供必要的人工智能研发技能（见附文）。这些培训机会针对的是为AI研发创新做出贡献的科学家和工程师以及可能拥有相关领域知识的AI研发用户。就前者而言，如战略1所述，联邦对人工智能研发的长期投资进一步支持了这一劳动力的增长，既培养了下一代研究人员，又使教师职位对当前的研究生和博士后更具吸引力。学生们。就后者而言，新计划正在为人工智能系统的当前和未来用户带来与人工智能相关的技能（见附文）。因此，联邦机构必须继续战略性地培养跨越多个学科和技能类别的人工智能研发人员的专业知识，以确保持续的国家领导。

---

[118] https://github.com/touretzkyds/ai4k12/wiki

[119] https://www.nap.edu/catalog/24926/assessing-and-responding-to-the-growth-of-computer-science- 本科招生

[120] https://www.whitehouse.gov/wp-content/uploads/2018/12/STEM-Education-Strategic-Plan-2018.pdf

实现本战略中概述的所需AI研发进展将需要足够的AI研发人员。人工智能研发领域最强大的国家将在未来的自动化领域建立领先地位。他们将成为算法创建和开发等能力的领跑者;能力论证;和商业化。发展技术专长将为这些进步提供基础。

虽然目前没有正式的人工智能劳动力数据，但商业和学术部门最近的大量报告表明，人工智能现有专家的数量不断增加。据报道AI专家供不应求，[121] 需求预计将继续升级。[122] 据报道，高科技公司正在投入大量资源来招聘具有AI专业知识的教师和学生。[123] 据报道，大学和工业正在招募和留住人才。[124]

需要进一步的研究，以更好地了解人工智能研发的当前和未来国家劳动力需求。需要数据来描述AI研发人员的当前状态，包括学术界，政府和行业的需求。研究应该探索AI工作场所的供需力量，以帮助预测未来的劳动力需求。需要了解预计的AI研发劳动力管道。应包括对教育途径和潜在再培训机会的考虑。还应探索多样性问题，因为研究表明，多样化的信息技术劳动力可以带来改善的结果。[125] 一旦更好地理解当前和未来的AI研发人员需求，就可以考虑采取适当的计划和行动来解决任何现有或预期的劳动力挑战。

[121]"初创公司旨在利用深度学习技能差距"，麻省理工学院技术评论，2016年1月6日。

[122]"人才攫取引发兴奋和担忧，"大自然，2016年4月26日。

[123]"人工智能专家需求量大"，"华尔街日报"，2015年5月1日。

[124]"百万美元的婴儿：随着硅谷争夺人才，大学努力抓住他们的明星，"经济学人，2016年4月2日。

[125]JW Moody，CM Beise，AB Woszczynski和ME Myers，"多样性与信息技术劳动力：障碍与机遇"，计算机信息系统期刊43（2003）：63-71。

战略8：扩大公私合作伙伴关系，加速人工智能的发展

战略8在2019年是新的，反映了实现人工智能研发的公私伙伴关系日益增长的重要性。

美国在科学和工程研究与创新方面的领导地位植根于国家独特的政府 – 大学 – 产业研发生态系统。正如美国艺术与科学协会所写，“美国作为创新领导者的地位”依赖于“建立一个更强大的国家政府 – 大学 – 工业研究伙伴关系。“[126] 自2016年国家人工智能研发战略计划发布以来，政府已经扩大了这一愿景，即与学术界，工业界，国际合作伙伴和盟国以及其他非联邦实体合作，推动“对人工智能研发的持续投资，以实现人工智能的技术突破和相关技术，并将这些突破迅速转变为有助于美国经济和国家安全的能力。“[1]

在过去的几十年里，在拥有联邦资金和工业的大学进行的信息技术基础研究已经为国家经济带来了新的，数十亿美元的新领域。[127] 政府，大学和工业界的共同进步是相辅相成的，并导致了一个创新，充满活力的人工智能部门。
今天的许多人工智能系统都是如此
由美国政府 – 大学 – 工业研发生态系统实现（见附文）。

自2016年国家人工智能研发战略计划发布以来，人们更加重视公私伙伴关系的益处。这些好处包括战略性地利用资源，包括设施，数据集和专业知识，以推动科学和工程创新；

推进国家的人工智能创新生态系统，涵盖政府，大学和行业

- 深度卷积神经网络已被证明是植根于人工智能研究的关键创新。虽然这种建模方法源自20世纪80年代后期的早期联邦投资，但当时神经网络没有足够的数据和足够的计算能力来进行准确的预测。通过大数据的增加，当今数据密集型科学方法的结合，以及如何构建和优化网络的概念上的进步，神经网络已经重新成为提高AI模型准确性的有效方法。近年来，学术界和私营部门之间的相互作用，包括政府资助，有助于降低语音识别系统的错误率，实现实时翻译等创新。.[126]
- 同样，联邦在20世纪80年代和90年代对强化学习的投资 – 一种植根于行为心理学的方法，涉及学习将行为与期望的结果联系起来 – 导致了今天的深度学习系统。通过跨部门的互动，计算机越来越像人类一样学习，没有明确的指导，强化学习推动了自动驾驶汽车和其他形式的自动化的进步，其中机器可以通过经验磨练技能。强化学习是AlphaGo的关键技术，该计划击败了世界上最好的围棋选手，自2016年以来，该选手已经取得了越来越多的职业选手胜利。[126]

[126] *恢复基础：研究在保护美国梦中的重要作用（美国艺术与科学学院，剑桥，马萨诸塞州，2014年）*；https://www.amacad.org/multimedia/pdfs/publications/researchpapersmonographs 美国基础研究基金会.

[127] 国家研究委员会计算机科学电信委员会，信息技术的持续创新（国家科学院出版社，华盛顿特区，2012年）；https://doi.org/10.17226/13427.

加速这些创新向实践的转变；并加强对下一代研究人员，技术人员和领导者的教育和培训。政府 – 大学 – 产业研发合作伙伴关系为大学研究人员带来了工业所面临的紧迫的现实挑战，实现了“使用启发式研究”；利用行业专业知识，加速将开放和已发表的研究成果转化为市场中可行的产品和服务，促进经济增长；通过将大学教师和学生与行业代表，行业环境和行业工作联系起来，增加研究和劳动力的能力（见附文）。[126,128,129,130] 这些伙伴关系建立在联邦机构之间的联合参与的基础上，在机构任务交叉的领域实现协同增效。国家也受益于联邦机构和国际资助者之间的关系，他们可以共同努力解决各个学科共同关心的关键挑战。

虽然公私伙伴关系有许多结构和机制，但参与的一些常见类别包括：

1. *基*于项目的单独协作。这些努力使大学研究人员能够与其他部门（包括联邦机构，行业和国际组织）的人员进行接触，以确定和利用共同感兴趣领域的协同作用。
2. *促*进开放，竞争前的基础研究的联合计划。跨部门组织之间的直接合作伙伴关系可以为合作伙伴共同感兴趣的领域提供资金和支持开放，竞争前的基础研究。一般而言，提供研究资源的非联邦合作伙伴获得Bayh-Dole法案赋予美国政府相同的知识产权。[131]
3. *合*作部署和加强研究基础设施。联邦机构，行业和国际组织之间的合作显着增强了开发新的和增强现有研究基础设施的潜力，从而使研究人员能够进行开创性的实验。合作伙伴可以提供财务和/或实物支持，以开发和/或增强研究基础设施。
4. *合*作以加强劳动力发展，包括扩大参与。多部门合作伙伴关系为严谨，有吸引力和鼓舞人心的教学材料奠定了基础，这些材料可以增强STEM专业的劳动力发展和多样性。

在每一种情况下，为了所有人的利益而利用每个合作伙伴的优势对于取得成功至关重要。

---

[128]数*学科学研究所报告，“伙伴关系：NSF ／ MPS与私人基金会合作研讨会”，2015年*；http://library.msri.org/msri/Partnerships.pdf.

[129]计*算社区联盟，“计算研究的未来：产业 – 学术合作”，2016年*；http://cra.org/ccc/wp-content/uploads/sites/2/2016/06/15125-CCC-Industry-Whitepaper-v4-1.pdf.

[130]计*算社区联盟，“计算研究中不断发展的学术/行业关系：CCC发布的中期报告”，2019年*；https://www.cccblog.org/wp-content/uploads/2019/03/Industry-Interim-Report-w- footnotes.pdf。

[131] https://history.nih.gov/research/downloads/PL96-517.pdf

人工智能研发的进步将受益于所有这些类型的公共私营合作伙伴关系。伙伴关系可以促进开放的，竞争前的，基础的人工智能研发；增强对数据集，模型和高级计算能力等研究资源的访问；并促进政府，大学和行业之间的研究人员交流和/或联合任命，以分享AI研发专业知识。伙伴关系也可以发布该人工智能研发具有固有的跨学科性质，需要计算机和信息科学，认知科学和心理学，经济学和博弈论，工程和控制理论，伦理学，语言学，数学和统计学以及哲学之间的融合，以推动未来AI的发展和评估系统公平，透明，负责，安全可靠。联邦机构正在积极寻求公私合作以实现这些目标（见附文）。

因此，联邦机构必须继续追求和加强人工智能研发中的公私合作伙伴关系，通过在政府，行业和学术界共同关心的领域中利用投资和专业知识来推动技术开发和经济增长。在这样做的过程中，美国政府将利用一个独特的美国创新生态系统，通过新颖的信息技术在过去二十年中改变了国家经济和社会的各个方面。.[127]

公私合作伙伴关系：近期的机构研发计划

许多机构已经启动了公私合作伙伴关系，以支持人工智能研发：

- 国防创新部门（DIU）[132] 是一个国防部组织，寻求能够满足国防部需求的商业解决方案。DIU反过来提供试验合同，其中可包括硬件，软件或其他独特服务。如果成功，试点合同将导致公司与任何国防部实体之间的后续合同。DIU的一个关键特征是试点和后续合同的快速发展。
- NSF和人工智能合作伙伴关系是一个多元化，多利益相关方组织，旨在更好地了解AI的影响，正在合作共同支持AI的社会和技术层面的高风险，高回报研究.[15]
- 美国国土安全部科技理事会硅谷创新计划（SVIP）[133] 希望利用全国乃至全球的商业研发创新生态系统来实现政府应用的技术。SVIP采用简化的应用程序和音调流程；将政府，企业家和行业聚集在一起，寻找最前沿的解决方案；并共同投资并加速向市场的过渡。
- 卫生和人类服务部（HHS）试行了Health Tech Sprint计划，该计划在其第一次迭代中称为"Top Health"，部分模仿人口普查局的机会项目。这项工作为围绕双向数据链路的公私合作创建了一个灵活的框架。它试验了用于迭代数据发布的新模型，用于人工智能培训和测试，并为公共 - 私人AI生态系统开发了一个自愿的激励框架。
- HHS研究，创新和风险投资部门是负责准备和响应助理秘书办公室生物医学高级研究和发展管理局的一部分。它负责监督加速器网络并正在招募一个非营利性合作伙伴，该合作伙伴可以与私人投资者合作，为创新技术和产品提供资金，以解决系统性健康安全挑战，人工智能应用程序是其中一个受关注的领域。加速器将连接初创公司和其他企业的产品开发和业务支持服务。

[132] https://www.diu.mil/

[133] https://www.dhs.gov/science-and-technology/svip

缩略语

空军科学研究局 空军科学研究办公室
嗳 人工智能
达帕 国防高级研究计划局
美国国土安全部 国土安全部
国防部 国防部
母鹿 能源部
点 运输部FDA 食品药品管理局FRVT 人脸识别供应商测试GPS 全球定位系统
图形处理器 图形处理单元
GSA 总务管理局
卫生部 卫生与人类服务部
高性能混凝土 高性能计算
亚尔帕 情报高级研究项目活动
IEC 国际电工委员会
IEEE 电气和电子工程师协会
碰撞 网络风险与信任政策与分析信息市场（DHS）
异 国际标准化组织
它 信息技术
免疫球蛋白组 机构间工作组
毫升 机器学习
麦莱 机器学习和人工智能（NSTC小组委员会）
美国航空航天局 国家航空和航天局
军士 NITRD国家协调办公室
NDS 自然驾驶研究（DOT）
的NifA 国家粮食和农业研究所（USDA）
美国国立卫生研究院 国立卫生研究院
NIST 国家标准与技术研究所
硝石 网络和信息技术研究与发展计划
神经网络模型 国家医学图书馆（NIH）
NSF 国家科学基金会NSTC 国家科技评议会
国家航空航天局 国家电信和信息管理局
奥德尼 国家情报局局长办公室
奥斯特普 科技政策办公室
R&D 研究与开发
射频干扰 索取资料
干 科学，技术，工程和数学
SVIP 硅谷创新计划（DHS）
特雷克 文本检索会议
美国农业部 美国农业部
佤族 美国退伍军人事务部
赛 可解释的AI

国家科学技术委员会

椅子

OSTP主任Kelvin Droegemeier

员工

ChloéKontos，NSTC执行董事

人工智能专责委员会

联合主席

Michael Kratsios，技术政策总裁助理（白宫）

法国A.Córdova，NSF主任

DARPA主任Steven Walker

机器学习和人工智能小组委员会

联合主席

Lynnet Parker，OSTP人工智能助理总监

美国国家科学基金会计算机信息科学与工程理事会（CISE）助理主任James Kurose

NIST信息技术实验室主任Charles Romine

美国能源部科学办公室科学项目副主任Stephen Binkley

行政秘书

费萨尔-德苏扎

网络与信息技术研究与发展小组委员会

联合主席

Kamie Roberts，NITRD NCO主任

美国国家科学基金会CISE助理总监James Kurose

行政秘书

Nekeia Butler

人工智能研究与开发机构间工作组

联合主席

IARPA国家情报总监（ODNI）项目经理Jeff Alstott

美国国家科学基金会信息与智能系统CISE部门主任Henry Kautz

员工

NITRD NCO技术协调员Faisal D'Souza

金伯利弗格森 - 沃尔特，美国海军

战略计划编写团队

Jeff Alstott，IARPA Gil Alterovitz，VA

Sameer Antani，美国国立卫生研究院

夏洛特贝尔，NIFA，美国农业部

Daniel Clouse

费萨尔-德苏扎

Michael Garris，NIST欧文

GiangdANi，NSF罗斯吉尔菲兰，OSTP特拉维斯厅，NTIA

美国国家科学基金会梅根霍顿

Henry Kautz，NSF

Erink NeNLY，DHS

David Kuehn，点

James Kurose，
NSF杰姆斯劳顿，
AFOSR Steven
Lee，DOE Aaron
Mannes，DHS琳恩
派克，OSTP
Dinesh Patwardhan
NIST的Elham Tabassi

关于国家科学技术委员会

国家科学技术委员会（NSTC）是行政部门在构成联邦研究和开发企业的各种实体之间协调科学和技术政策的主要手段。NSTC的主要目标是确保科学和技术政策决策和计划符合总统的既定目标。NSTC准备了旨在实现多个国家目标的联邦机构协调的研究和发展战略。NSTC的工作由委员会组织，委员会负责监督关注科学和技术不同方面的小组委员会和工作组。有关更多信息，请访问：https://www.whitehouse.gov/ostp/nstc.

关于科技政策办公室

科学和技术政策办公室（OSTP）是根据1976年国家科学和技术政策，组织和优先事项法案设立的，为总统和总统执行办公室内的其他人提供科学，工程和技术方面的建议。经济，国家安全，国土安全，健康，对外关系，环境，技术恢复和资源利用等方面。OSTP领导机构间科学和技术政策协调工作，协助管理和预算办公室对预算中的联邦研究与开发（R&D）进行年度审查和分析，并作为总统的科学和技术分析和判断来源。尊重联邦政府的主要政策，计划和计划。有关更多信息，请访问：https://www.whitehouse.gov/ostp.

关于人工智能专责委员会

人工智能特别委员会（AI）建议并协助NSTC提高与AI相关的联邦研发工作的整体效率和生产力，以确保美国在该领域的持续领导地位。它涉及跨越机构界限的国家和国际政策事务，并为联邦人工智能研发活动提供机构间政策协调和发展的正式机制，包括与自治系统，生物识别，计算机视觉，人机交互，机器学习，自然相关的活动。语言处理和机器人技术。它还就机构间人工智能研发优先事项向总统执行办公室提供建议;致力于创建平衡和全面的AI研发计划和合作伙伴关系;利用跨部门和机构任务的联邦数据和计算资源;并支持技术，国家AI人力资源。

关于机器学习和人工智能小组委员会

机器学习和人工智能（MLAI）小组委员会监督联邦政府，私营部门和国际机器学习（ML）和人工智能的最新技术，以监测重要技术里程碑的发展。 AI，协调联邦政府使用和促进有关ML和AI的知识和最佳实践的分享，并参与制定联邦MLAI研发优先事项。MLAI小组委员会向技术委员会和AI专门委员会报告。MLAI小组委员会还与人工智能研究与发展机构间工作组（见下文）协调AI任务。

关于网络与信息技术研究与发展小组委员会

网络和信息技术研究与发展（NITRD）计划是Nation在联邦政府资助的计算，网络和软件领域的先驱信息技术（IT）工作的主要来源。NITRD小组委员会指导多机构NITRD计划的工作，为确保美国技术领先地位和满足国家对先进IT的需求提供研发基础。它向NSTC科学技术企业委员会报告。小组委员会得到向其及其国家协调办公室报告的机构间工作组的支持。有关更多信息，请访问：https://www.nitrd.gov/about/.

关于人工智能研究与开发跨部门工作组

NITRD人工智能研发机构间工作组（IWG）协调人工智能的联邦研发;它还支持和协调人工智能专题委员会和NSTC机器学习和人工智能小组委员会的任务。这项重要工作促进了美国在人工智能研发方面的领导地位和全球竞NITRD人工智能研发IWG牵头更新了这项国家人工智能研究与发展战略计划。有关更多信息，请访问：https://www.nitrd.gov/groups/AI.

关于本文件

本文件包括2016年国家AI研发战略计划的原始文本，其中包括2016年计划管理和机构间评估后的2019年更新以及社区对更新计划信息请求的回复。2016年的战略被广泛认为是有效的，并且需要进行一些重新审核，并呼吁制定关于人工智能的私人 – 公共伙伴关系的新战略。每个策略的顶部都插入了一个带阴影的标注框，以突出显示更新的命令和/或新的焦点区域。2019年的更新增加了关于人工智能私人 – 公共伙伴关系的全新战略8。

版权信息

发表于2019年的美利坚合众国。

THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN: 2019 UPDATE

A Report by the

SELECT COMMITTEE ON ARTIFICIAL INTELLIGENCE

of the

NATIONAL SCIENCE & TECHNOLOGY COUNCIL

JUNE 2019

Dear Colleagues,

In his State of the Union address on February 5, 2019, President Trump stressed the importance of ensuring American leadership in the development of emerging technologies, including artificial intelligence (AI), that make up the Industries of the Future. Reflecting this importance, on February 11, 2019, President Trump signed Executive Order 13859, which established the American Artificial Intelligence Initiative. This Initiative is a whole-of-government approach for maintaining American leadership in AI and ensuring that AI benefits the American people and reflects our Nation's values. The first directive in this Executive Order is for Federal agencies to prioritize AI research and development (R&D) in their annual budgeting and planning process. The attached *National AI R&D Strategic Plan: 2019 Update* highlights the key priorities for Federal investment in AI R&D.

Artificial intelligence presents tremendous opportunities that are leading to breakthroughs in improved healthcare, safer and more efficient transportation, personalized education, significant scientific discoveries, improved manufacturing, increased agricultural crop yields, better weather forecasting, and much more. These benefits are largely due to decades of long-term Federal investments in fundamental AI R&D, which have led to new theories and approaches for AI systems, as well as applied research that allows the translation of AI into practical applications.

The landscape for AI R&D is becoming increasingly complex, due to the significant investments that are being made by industry, academia, and nonprofit organizations. Additionally, AI advancements are progressing rapidly. The Federal Government must therefore continually reevaluate its priorities for AI R&D investments, to ensure that investments continue to advance the cutting edge of the field and are not unnecessarily duplicative of industry investments.

In August of 2018, the Administration directed the Select Committee on AI to refresh the 2016 *National AI R&D Strategic Plan*. This process began with the issuance of a Request for Information to solicit public input on ways that the strategy should be revised or improved. The responses to this RFI, as well as an independent agency review, informed this update to the Strategic Plan.

In this Strategic Plan, eight strategic priorities have been identified. The first seven strategies continue from the 2016 Plan, reflecting the reaffirmation of the importance of these strategies by multiple respondents from the public and government, with no calls to remove any of the strategies. The eighth strategy is new and focuses on the increasing importance of effective partnerships between the Federal Government and academia, industry, other non-Federal entities, and international allies to generate technological breakthroughs in AI and to rapidly transition those breakthroughs into capabilities.

While this Plan does not define specific research agendas for Federal agency investments, it does provide an expectation for the overall portfolio for Federal AI R&D investments. This coordinated Federal strategy for AI R&D will help the United States continue to lead the world in cutting-edge advances in AI that will grow our economy, increase our national security, and improve quality of life.

Sincerely,

Michael Kratsios
Deputy Assistant to the President for Technology Policy
June 21, 2019

Table of Contents

Executive Summary

Artificial intelligence (AI) holds tremendous promise to benefit nearly all aspects of society, including the economy, healthcare, security, the law, transportation, even technology itself. On February 11, 2019, the President signed Executive Order 13859, *Maintaining American Leadership in Artificial Intelligence*.[1] This order launched the American AI Initiative, a concerted effort to promote and protect AI technology and innovation in the United States. The Initiative implements a whole-of-government strategy in collaboration and engagement with the private sector, academia, the public, and like-minded international partners. Among other actions, key directives in the Initiative call for Federal agencies to prioritize AI research and development (R&D) investments, enhance access to high-quality cyberinfrastructure and data, ensure that the Nation leads in the development of technical standards for AI, and provide education and training opportunities to prepare the American workforce for the new era of AI.

In support of the American AI Initiative, this *National AI R&D Strategic Plan: 2019 Update* defines the priority areas for Federal investments in AI R&D. This 2019 update builds upon the first *National AI R&D Strategic Plan* released in 2016, accounting for new research, technical innovations, and other considerations that have emerged over the past three years. This update has been developed by leading AI researchers and research administrators from across the Federal Government, with input from the broader civil society, including from many of America's leading academic research institutions, nonprofit organizations, and private sector technology companies. Feedback from these key stakeholders affirmed the continued relevance of each part of the 2016 Strategic Plan while also calling for greater attention to making AI trustworthy, to partnering with the private sector, and other imperatives.

The *National AI R&D Strategic Plan: 2019 Update* establishes a set of objectives for Federally funded AI research, identifying the following eight strategic priorities:

Strategy 1: Make long-term investments in AI research. Prioritize investments in the next generation of AI that will drive discovery and insight and enable the United States to remain a world leader in AI.

Strategy 2: Develop effective methods for human-AI collaboration. Increase understanding of how to create AI systems that effectively complement and augment human capabilities.

Strategy 3: Understand and address the ethical, legal, and societal implications of AI. Research AI systems that incorporate ethical, legal, and societal concerns through technical mechanisms.

Strategy 4: Ensure the safety and security of AI systems. Advance knowledge of how to design AI systems that are reliable, dependable, safe, and trustworthy.

Strategy 5: Develop shared public datasets and environments for AI training and testing. Develop and enable access to high-quality datasets and environments, as well as to testing and training resources.

Strategy 6: Measure and evaluate AI technologies through standards and benchmarks. Develop a broad spectrum of evaluative techniques for AI, including technical standards and benchmarks.

Strategy 7: Better understand the national AI R&D workforce needs. Improve opportunities for R&D workforce development to strategically foster an AI-ready workforce.

Strategy 8: Expand public-private partnerships to accelerate advances in AI. Promote opportunities for sustained investment in AI R&D and for transitioning advances into practical capabilities, in collaboration with academia, industry, international partners, and other non-Federal entities.

[1] https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

Introduction to the 2019 National AI R&D Strategic Plan

Artificial intelligence enables computers and other automated systems to perform tasks that have historically required human cognition and what we typically consider human decision-making abilities. Over the past several decades, AI has advanced tremendously and today promises better, more accurate healthcare; enhanced national security; improved transportation; and more effective education, to name just a few benefits. Increased computing power, the availability of large datasets and streaming data, and algorithmic advances in machine learning (ML) have made it possible for AI development to create new sectors of the economy and revitalize industries. As more industries adopt AI's fundamental technologies, the field will continue to drive profound economic impact and quality-of-life improvements worldwide.

These advancements have been driven primarily by Federal investments in AI R&D, the expertise of America's unsurpassed R&D institutions, and the collective creativity of many of America's most visionary technology companies and entrepreneurs.

In 2016 the Federal Government published first *National AI R&D Strategic Plan*, recognizing AI's tremendous promise and need for continued advancement. It was developed to guide the Nation in our AI R&D investments, provide a strategic framework for improving and leveraging America's AI capabilities, and ensure that those capabilities produce prosperity, security, and improved quality of life for the American people for years to come.

The Plan defined several key areas of priority focus for the Federal agencies that invest in AI. These focus areas, or strategies, include: continued long-term investments in AI; effective methods for human-AI collaboration; understanding and addressing the ethical, legal, and societal implications for AI; ensuring the safety and security of AI; developing shared public datasets and environments for AI training and testing; measuring and evaluating AI technologies through standards and benchmarks; and better understanding the Nation's AI R&D

2019 Update	**RFI responses inform the 2019 National AI R&D Strategic Plan**

In September 2018, the National Coordination Office for Networking and Information Technology Research and Development issued a Request for Information (RFI)[2] on behalf of the Select Committee on Artificial Intelligence, requesting input from all interested parties on the 2016 *National Artificial Intelligence Research and Development Strategic Plan*. Nearly 50 responses were submitted by researchers, research organizations, professional societies, civil society organizations, and individuals; these responses are available online.[3]

Many of the responses reaffirmed the analysis, organization, and approach outlined in the 2016 *National AI R&D Strategic Plan*. A significant number of responses noted the importance of investing in the application of AI in areas such as manufacturing and supply chains; healthcare; medical imaging; meteorology, hydrology, climatology, and related areas; cybersecurity; education; data-intensive physical sciences such as high-energy physics; and transportation. This interest in translational applications of AI technologies has certainly increased since the release of the 2016 *National AI R&D Strategic Plan*. Other common themes echoed in the RFI responses were the importance of developing trustworthy AI systems, including fairness, ethics, accountability, and transparency of AI systems; curated and accessible datasets; workforce considerations; and public-private partnerships for furthering AI R&D.

[2] https://www.nitrd.gov/news/RFI-National-AI-Strategic-Plan.aspx

[3] https://www.nitrd.gov/nitrdgroups/index.php?title=AI-RFI-Responses-2018

workforce needs. That work was prescient: today, countries around the world have followed suit and have issued their own versions of this plan.

In the three years since the *National AI R&D Strategic Plan* was produced, new research, technical innovations, and real-world deployments have progressed rapidly. The Administration initiated this 2019 update to the *National AI R&D Strategic Plan* to address these advancements, including a rapidly evolving international AI landscape.

Notably, this 2019 Update to the *National AI R&D Strategic Plan* is, by design, solely concerned with addressing the *research and development* priorities associated with advancing AI technologies. It does not describe or recommend policy or regulatory actions related to the governance or deployment of AI, although AI R&D will certainly inform the development of reasonable policy and regulatory frameworks.

AI as an Administration Priority

Since 2017, the Administration has addressed the importance of AI R&D by emphasizing its role for America's future across multiple major policy documents, including the *National Security Strategy*,[4] the *National Defense Strategy*,[5] and the FY 2020 R&D Budget Priorities Memo.[6]

In May 2018, the Office of Science and Technology Policy (OSTP) hosted the White House Summit on Artificial Intelligence for American Industry to begin discussing the promise of AI and the policies needed to realize that promise for the American people and maintain U.S. leadership in the age of AI. The Summit convened over 100 senior government officials, technical experts from top academic institutions, heads of industrial research laboratories, and American business leaders.

In his State of the Union address on February 5, 2019, President Trump stressed the importance of ensuring American leadership in the development of emerging technologies, including AI, that make up the Industries of the Future.

On February 11, 2019, the President signed Executive Order 13859, *Maintaining American Leadership in Artificial Intelligence*.[7] This order launched the American AI Initiative, a concerted effort to promote and protect AI technology and innovation in the United States. The Initiative implements a whole-of-government strategy in collaboration and engagement with the private sector, academia, the public, and like-minded international partners. Among other actions, key directives in the Initiative call for Federal agencies to prioritize AI R&D investments, enhance access to high-quality cyberinfrastructure and data, ensure that the Nation leads in the development of technical standards for AI, and provide education and training opportunities to prepare the American workforce for the new era of AI.

Development of the 2019 Update to the *National AI R&D Strategic Plan*

The 2016 *National AI R&D Strategic Plan* recommended that the many Federal agencies tasked with advancing or adopting AI collaborate to identify critical R&D opportunities and support effective coordination of Federal AI R&D activities, both intramural and extramural research. Reflecting the Administration's prioritization of AI, the National Science and Technology Council (NSTC) has established a new framework to implement this recommendation, consisting of three unique NSTC subgroups made up of members from across the Federal R&D agencies to cover (1) senior leadership

[4] https://www.whitehouse.gov/wp-content/uploads/2017/12/NSS-Final-12-18-2017-0905.pdf

[5] https://dod.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf

[6] https://www.whitehouse.gov/wp-content/uploads/2018/07/M-18-22.pdf

[7] https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

and strategic vision, (2) operational planning and tactical implementation, and (3) research and technical expertise. These subgroups are:

- The Select Committee on AI,[8] consisting of the heads of departments and agencies principally responsible for the government's AI R&D, advises the Administration on interagency AI R&D priorities; considers the creation of Federal partnerships with industry and academia; establishes structures to improve government planning and coordination of AI R&D; identifies opportunities to leverage Federal data and computational resources to support our national AI R&D ecosystem; and supports the growth of a technical, national AI workforce.
- The NSTC Subcommittee on Machine Learning and Artificial Intelligence (MLAI), consisting of agency AI leaders and administrators, serves as the operational and implementation arm of the Select Committee, responsible for fulfilling tasking from the Select Committee; creating and maintaining the *National AI R&D Strategic Plan*; identifying and addressing important policy issues related to AI research, testing, standards, education, implementation, outreach, and related areas; and related activities.
- The AI R&D Interagency Working Group, operating under the NSTC's Networking and Information Technology R&D (NITRD) Subcommittee and consisting of research program managers and technical experts from across the Federal Government, reports to the MLAI Subcommittee; helps coordinate interagency AI R&D programmatic efforts; serves as the interagency AI R&D community of practice; and reports government-wide AI R&D spending through the NITRD Subcommittee's annual Supplement to the President's Budget.

In September 2018, the Select Committee initiated an update to the 2016 Strategic Plan, beginning with an RFI seeking broad community input on whether and how the seven strategies of the 2016 *National AI R&D Strategic Plan* merited revision or replacement (see sidebar). Independently, Federal departments and agencies performing or funding AI R&D undertook their own assessments.

An Overview of the 2019 Update to the 2016 *National AI R&D Strategic Plan*

Together, the Select Committee on AI, the NSTC Subcommittee on Machine Learning and AI, and the AI R&D Interagency Working Group of NITRD reviewed the input regarding the *National AI R&D Strategic Plan.* Each of the original seven focus areas or strategies of the 2016 Plan was reaffirmed by multiple respondents from the public and government, with no calls to remove any one strategy. These strategies, updated in this 2019 Update to the Strategic Plan to reflect the current state of the art, are:

Strategy 1: Make long-term investments in AI research;

Strategy 2: Develop effective methods for human-AI collaboration;

Strategy 3: Understand and address the ethical, legal, and societal implications of AI;

Strategy 4: Ensure the safety and security of AI systems;

Strategy 5: Develop shared public datasets and environments for AI training and testing;

Strategy 6: Measure and evaluate AI technologies through standards and benchmarks; and

Strategy 7: Better understand the national AI R&D workforce needs.

[8] https://www.whitehouse.gov/wp-content/uploads/2018/05/Summary-Report-of-White-House-AI-Summit.pdf

Many responses to the RFI called for greater Federal Government R&D engagement with the private sector, given the fast rise of privately funded AI R&D, and the rapid adoption of AI by industry. As a result, the 2019 Update incorporates a new, eighth strategy:

Strategy 8: Expand public-private partnerships to accelerate advances in AI.

Feedback from the public and Federal agencies identified a number of specific challenges to further AI development and adoption. These challenges, many of which cut across multiple agencies, provide enhanced insight into ways that this *National AI R&D Strategic Plan* can guide the course of AI R&D in America, and many closely relate to the themes addressed in the 2019 *Executive Order on Maintaining American Leadership in Artificial Intelligence.* Examples include the following:

- *Research at the frontiers.* Even though machine learning has brought phenomenal new capabilities in the past several years, continued research is needed to further push the frontiers of ML, as well as to develop additional approaches to the tough technical challenges of AI (Strategy 1).
- *Positive impact.* As AI capabilities grow, the United States must place increased emphasis on developing new methods to ensure that AI's impacts are robustly positive into the future (Strategies 1, 3, and 4).
- *Trust and explainability.* Truly trustworthy AI requires explainable AI, especially as AI systems grow in scale and complexity; this requires a comprehensive understanding of the AI system by the human user and the human designer (Strategies 1, 2, 3, 4, and 6).
- *Safety and security*. Researchers must devise methods to keep AI systems and the data they use secure so that the Nation can leverage the opportunities afforded by this technology while also maintaining confidentiality and safety (Strategies 4, 5, and 6).
- *Technical standards.* As the Nation develops techniques to expand both AI abilities and assurance, it must test and benchmark them; when the techniques are ready, they should be turned into technical standards for the world (Strategy 6).
- *Workforce capability.* Accomplishing these goals will require growing a skilled AI R&D workforce that is currently limited and in high demand; the United States must be creative and bold in training and acquiring the skilled workforce it needs to lead the world in AI research and applications (Strategy 7).
- *Partnerships.* Advances in AI R&D increasingly require effective partnerships between the Federal Government and academia, industry, and other non-Federal entities to generate technological breakthroughs in AI and to rapidly transition those breakthroughs into capabilities (Strategy 8).
- *Cooperation with allies.* Additionally, the Plan recognizes the importance of international cooperation for successful implementation of these goals, while protecting the American AI R&D enterprise from strategic competitors and adversarial nations.

Structure of this 2019 Update to the 2016 *National AI R&D Strategic Plan*

This updated *National AI R&D Strategic Plan* incorporates the original text from the 2016 version, including the following section on R&D Strategy (except for minor edits) and the original 2016 wording of the first seven strategies. For each strategy, *2019 updates to the 2016 National R&D Strategic Plan are provided in shaded boxes at the top of the original seven strategies; these highlight updated imperatives and/or new focus areas for the strategies.* Text below the shaded boxes is *as it originally appeared* in the 2016 *National AI R&D Strategic Plan*, providing observations and context that remain important today (note that some of the original details may have become out of date in the intervening period). In addition, as noted previously, a new eighth strategy is added in this 2019 Update, on expanding public-private partnerships in AI R&D.

AI R&D Strategy

The research priorities outlined in this AI R&D Strategic Plan focus on areas that industry is unlikely to address on their own, and thus, areas that are most likely to benefit from Federal investment. These priorities cut across all of AI to include needs common to the AI sub-fields of perception, automated reasoning/planning, cognitive systems, machine learning, natural language processing, robotics, and related fields. Because of the breadth of AI, these priorities span the entire field, rather than only focusing on individual research challenges specific to each sub-domain. To implement the plan, detailed roadmaps should be developed that address the capability gaps consistent with the plan.

One of the most important Federal research priorities, outlined in Strategy 1, is for sustained long-term research in AI to drive discovery and insight. Many of the investments by the U.S. Federal Government in high-risk, high-reward[9] fundamental research have led to revolutionary technological advances we depend on today, including the Internet, GPS, smartphone speech recognition, heart monitors, solar panels, advanced batteries, cancer therapies, and much, much more. The promise of AI touches nearly every aspect of society and has the potential for significant positive societal and economic benefits. Thus, to maintain a world leadership position in this area, the United States must focus its investments on high-priority fundamental and long-term AI research.

Many AI technologies will work with and alongside humans, thus leading to important challenges in how to best create AI systems that work with people in intuitive and helpful ways.[10] The walls between humans and AI systems are slowly beginning to erode, with AI systems augmenting and enhancing human capabilities. Fundamental research is needed to develop effective methods for human-AI interaction and collaboration, as outlined in Strategy 2.

AI advancements are providing many positive benefits to society and are increasing U.S. national competitiveness.[11] However, as with most transformative technologies, AI presents some societal risks in several areas, from jobs and the economy to safety, ethical, and legal questions. Thus, as AI science and technology develop, the Federal Government must also invest in research to better understand what the implications are for AI for all these realms, and to address these implications by developing AI systems that align with ethical, legal, and societal goals, as outlined in Strategy 3.

A critical gap in current AI technology is a lack of methodologies to ensure the safety and predictable performance of AI systems. Ensuring the safety of AI systems is a challenge because of the unusual complexity and evolving nature of these systems. Several research priorities address this safety challenge. First, Strategy 4 emphasizes the need for explainable and transparent systems that are trusted by their users, perform in a manner that is acceptable to the users, and can be guaranteed to act as the user intended. The potential capabilities and complexity of AI systems, combined with the wealth of possible interactions with human users and the environment, makes it critically important to invest in research that increases the security and control of AI technologies. Strategy 5 calls on the Federal Government to invest in shared public datasets for AI training and testing to advance the progress of AI research and to enable a more effective comparison of alternative solutions.

Strategy 6 discusses how standards and benchmarks can focus R&D to define progress, close gaps, and drive innovative solutions for specific problems and challenges. Standards and benchmarks are

[9] "High-risk, high-reward" research refers to visionary research that is intellectually challenging but has the potential to make deeply positive, transformative impacts on the field of study.

[10] See *2016 Report of the One Hundred Year Study on Artificial Intelligence*, which focuses on the anticipated uses and impacts of AI in the year 2030; https://ai100.stanford.edu/2016-report.

[11] J. Furman, "Is This Time Different? The Opportunities and Challenges of Artificial Intelligence," Council of Economic Advisors remarks, New York University: AI Now Symposium, July 7, 2016.

essential for measuring and evaluating AI systems and ensuring that AI technologies meet critical objectives for functionality and interoperability.

Finally, the growing prevalence of AI technologies across all sectors of society creates new pressures for AI R&D experts. Opportunities abound for core AI scientists and engineers with a deep understanding of the technology who can generate new ideas for advancing the boundaries of knowledge in the field. The Nation should take action to ensure a sufficient pipeline of AI-capable talent. Strategy 7 addresses this challenge.

Figure 1 (*updated in this 2019 version of the Plan*) provides a graphical illustration of the overall organization of this AI R&D Strategic Plan. Across the bottom row of boxes are the crosscutting, underlying foundations that affect the development of all AI systems; these foundations are described in Strategies 3-7 and the new Strategy 8. The next layer higher (middle row of boxes) includes many areas of research that are needed to advance AI. These R&D areas (including use-inspired basic research) are outlined in Strategies 1-2.[12] Across the top row of boxes in the graphic are examples of applications that are expected to benefit from advances in AI. Together, these components of the AI R&D Strategic Plan define a high-level framework for Federal investments that can lead to impactful advances in the field and positive societal benefits.

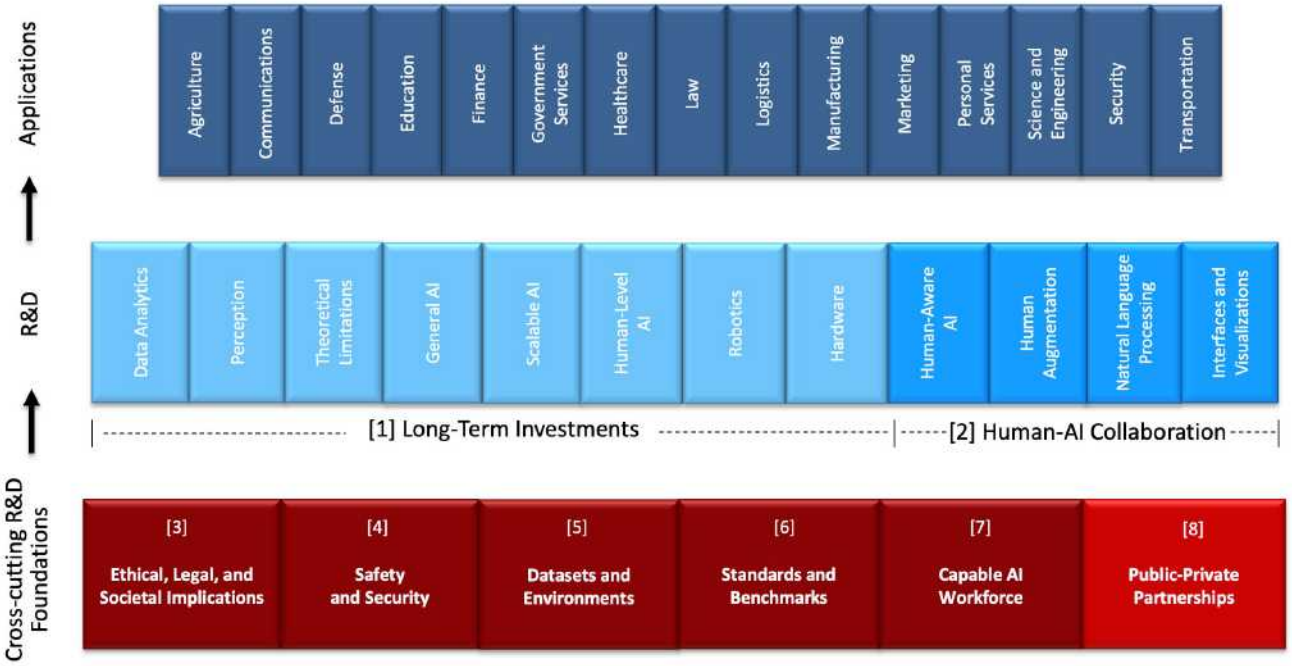

Figure 1. Organization of the AI R&D Strategic Plan (2019 update, to include Strategy 8). A combination of crosscutting R&D foundations (*in the lower row*) are important for all AI research. Many AI R&D areas (*in the middle row*) can build upon these crosscutting foundations to impact a wide array of societal applications (*in the top row*). The numbers in brackets indicate the number of the Strategy in this plan that further develops each topic. The ordering of these strategies does not indicate a priority of importance.

[12] Throughout this document, "basic research" includes both pure basic research and use-inspired basic research—the so-called Pasteur's Quadrant defined by Donald Stokes in his 1997 book of the same name—referring to basic research that has use for society in mind. For example, the fundamental NIH investments in IT are often called use-inspired basic research.

Strategy 1: Make Long-Term Investments in AI Research

2019 Update	**Sustaining long-term investments in fundamental AI research**

Since the release of the 2016 *National AI R&D Strategic Plan*, powerful new capabilities, primarily ML applications to well-defined tasks, have continued to emerge. These capabilities have demonstrated impacts in a diverse array of applications, such as classifying genetic sequences,[20,21] managing limited wireless spectrum resources,[22] interpreting medical images,[23] and grading cancers.[24] These rapid advances required decades of research for the technologies and applications to mature.[25] To maintain this progress in ML to achieve advancements in other areas of AI, and to strive toward the long-term goal of general-purpose AI, the Federal Government must continue to foster long-term, fundamental research in ML and AI. This research will give rise to transformational technologies and, in turn, breakthroughs across all sectors of society.

Much of the current progress in the field has been in specialized, well-defined tasks often driven by statistical ML, such as *classification*, *recognition*, and *regression* (i.e., "narrow AI systems"). Surveys of the

Long-term, fundamental AI research: Recent agency R&D programs

Since the release of the 2016 *National AI R&D Strategic Plan*, a number of agencies have initiated AI R&D programs for Strategy 1:

- NSF has continued to fund foundational research in AI, spanning ML, reasoning and representation, computer vision, computational neuroscience, speech and language, robotics, and multi-agent systems. NSF has launched new joint funding opportunities with other agencies—notably with DARPA in the area of high-performance, energy-efficient hardware for real-time ML[13] and with USDA-NIFA on AI for agricultural science[14]—and with industry.[15,16] In addition, NSF's Harnessing the Data Revolution Big Idea[17] supports research on the foundations of data science, which will serve as a driver of future ML and AI systems.
- DARPA announced in September 2018 a multiyear investment in new and existing programs called the "AI Next" campaign.[18] Key campaign areas include improving the robustness and reliability of AI systems; enhancing the security and resiliency of ML/AI technologies; reducing power, data, and performance inefficiencies; and pioneering the next generation of AI algorithms and applications, such as explainability and commonsense reasoning.
- The *NIH Strategic Plan for Data Science*[19] of September 2018 aims to advance access to data science technology and ML/AI capability for the biomedical research community toward data-driven healthcare research.

[13] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505640&org=NSF
[14] https://www.nsf.gov/pubs/2019/nsf19051/nsf19051.jsp
[15] https://www.nsf.gov/pubs/2019/nsf19018/nsf19018.jsp
[16] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505651
[17] https://www.nsf.gov/cise/harnessingdata/
[18] https://www.darpa.mil/work-with-us/ai-next-campaign
[19] https://datascience.nih.gov/strategicplan
[20] https://ai.googleblog.com/2017/12/deepvariant-highly-accurate-genomes.html
[21] https://irp.nih.gov/catalyst/v26i4/machine-learning
[22] https://www.spectrumcollaborationchallenge.com/
[23] https://news-medical.net/news/20190417/Workshop-explores-the-future-of-artificial-intelligence-in-medical-imaging.aspx
[24] https://www.nature.com/articles/nature21056
[25] https://www.nitrd.gov/rfi/ai/2018/AI-RFI-Response-2018-Yolanda-Gil-AAAI.pdf

field have noted that long-term investments in fundamental research are needed to continue building on these advances in ML. Further, parallel sustained efforts are required to fully realize the vision of "general-purpose AI"—systems that exhibit the flexibility and versatility of human intelligence in a broad range of cognitive domains.[26,27,28,29]

Emphasis is needed on the development of further ML capabilities to interactively and persistently learn, the connection between perception and attention, and the incorporation of learned models into comprehensive reasoning architectures.[30] Beyond ML, critical research is also needed in other core areas of AI, including in commonsense reasoning and problem solving, probabilistic reasoning, combinatorial optimization, knowledge representation, planning and scheduling, natural language processing, decision making, and human-machine interaction. Advances in these areas will in turn enable collaborative robotics and shared and fully autonomous systems (see Strategy 2). The grand challenge of understanding human intelligence requires significant investments in shared resources and infrastructure.[25] Broad consensus exists for foundational investments in drivers of ML and AI as well, including data provenance and quality, novel software and hardware paradigms and platforms, and the security of AI systems.[31,32] For example, as AI software performs increasingly complex functions in all aspects of daily life and all sectors of the economy, existing software development paradigms will need to evolve to meet software productivity, quality, and sustainability requirements.

Recent Federal investments have prioritized these areas of fundamental ML and AI research (see sidebar) as well as the use of ML and AI across numerous application sectors, including defense, security, energy, transportation, health, agriculture, and telecommunications. Ultimately, AI technologies are critical for addressing a range of long-term challenges, such as constructing advanced healthcare systems, a robust intelligent transportation system, and resilient energy and telecommunication networks.

For AI applications to become widespread, they must be explainable and understandable (see Strategy 3). These challenges are particularly salient for fostering collaborative human-AI relationships (see Strategy 2). Today, the ability to understand and analyze the decisions of AI systems and measure their accuracy, reliability, and reproducibility is limited. Sustained R&D investments are needed to advance trust in AI systems to ensure they meet society's needs and adequately address requirements for robustness, fairness, explainability, and security.

A long-term commitment to AI R&D is essential to continue and expand current technical advances and more broadly ensure that AI enriches the human experience. Indeed, the 2019 *Executive Order on Maintaining American Leadership in Artificial Intelligence* notes:

> Heads of implementing agencies that also perform or fund R&D (AI R&D agencies), shall consider AI as an agency R&D priority, as appropriate to their respective agencies' missions… Heads of such agencies shall take this priority into account when developing budget proposals and planning for the use of funds in Fiscal Year 2020 and in future years. Heads of these agencies shall also consider appropriate administrative actions to increase focus on AI for 2019.

[26] https://ai100.stanford.edu/sites/g/files/sbiybj9861/f/ai100report10032016fnl_singles.pdf
[27] http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf
[28] https://cra.org/ccc/visioning/visioning-activities/2018-activities/artificial-intelligence-roadmap/
[29] https://www.microsoft.com/en-us/research/research-area/artificial-intelligence/
[30] https://cra.org/ccc/events/artificial-intelligence-roadmap-workshop-3-learning-and-robotics/
[31] https://cra.org/ccc/wp-content/uploads/sites/2/2016/04/AI-for-Social-Good-Workshop-Report.pdf
[32] https://openai.com/blog/ai-and-compute/

AI research investments are needed in areas with potential long-term payoffs. While an important component of long-term research is incremental research with predictable outcomes, long-term sustained investments in high-risk research can lead to high-reward payoffs. These payoffs can be seen in 5 years, 10 years, or more. A 2012 National Research Council report emphasizes the critical role of Federal investments in long-term research, noting "the long, unpredictable incubation period—requiring steady work and funding—between initial exploration and commercial deployment."[33] It further notes that "the time from first concept to successful market is often measured in decades." Well-documented examples of sustained fundamental research efforts that led to high-reward payoffs include the World Wide Web and deep learning. In both cases, the basic foundations began in the 1960s; it was only after 30+ years of continued research efforts that these ideas materialized into the transformative technologies witnessed today in many categories of AI.

The following subsections highlight some of these areas. Additional categories of important AI research are discussed in Strategies 2 through 6.

Advancing data-focused methodologies for knowledge discovery

As discussed in the 2016 *Federal Big Data Research and Development Strategic Plan*,[34] many fundamental new tools and technologies are needed to achieve intelligent data understanding and knowledge discovery. Further progress is needed in the development of more advanced machine learning algorithms that can identify all the useful information hidden in big data. Many open research questions revolve around the creation and use of data, including its veracity and appropriateness for AI system training. The veracity of data is particularly challenging when dealing with vast amounts of data, making it difficult for humans to assess and extract knowledge from it. While much research has dealt with veracity through data quality assurance methods to perform data cleaning and knowledge discovery, further study is needed to improve the efficiency of data cleaning techniques, to create methods for discovering inconsistencies and anomalies in the data, and to develop approaches for incorporating human feedback. Researchers need to explore new methods to enable data and associated metadata to be mined simultaneously.

Many AI applications are interdisciplinary in nature and make use of heterogeneous data. Further investigation of multimodality machine learning is needed to enable knowledge discovery from a wide variety of different types of data (e.g., discrete, continuous, text, spatial, temporal, spatio-temporal, graphs). AI investigators must determine the amount of data needed for training and to properly address large-scale versus long-tail data needs. They must also determine how to identify and process rare events beyond purely statistical approaches; to work with knowledge sources (i.e., any type of information that explains the world, such as knowledge of the law of gravity or of social norms) as well as data sources, integrating models and ontologies in the learning process; and to obtain effective learning performance with little data when big data sources may not be available.

Enhancing the perceptual capabilities of AI systems

Perception is an intelligent system's window into the world. Perception begins with (possibly distributed) sensor data, which comes in diverse modalities and forms, such as the status of the system itself or information about the environment. Sensor data are processed and fused, often along with *a priori* knowledge and models, to extract information relevant to the AI system's task such as

[33] National Research Council Computer Science Telecommunications Board, *Continuing Innovation in Information Technology* (The National Academies Press, Washington, D.C., 2012), 11; https://doi.org/10.17226/13427.

[34] https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf

geometric features, attributes, location, and velocity. Integrated data from perception forms situational awareness to provide AI systems with the comprehensive knowledge and a model of the state of the world necessary to plan and execute tasks effectively and safely. AI systems would greatly benefit from advancements in hardware and algorithms to enable more robust and reliable perception. Sensors must be able to capture data at longer distances, with higher resolution, and in real time. Perception systems need to be able to integrate data from a variety of sensors and other sources, including the computational cloud, to determine what the AI system is currently perceiving and to allow the prediction of future states. Detection, classification, identification, and recognition of objects remain challenging, especially under cluttered and dynamic conditions. In addition, perception of humans must be greatly improved by using an appropriate combination of sensors and algorithms, so that AI systems can work more effectively with people.[10] A framework for calculating and propagating uncertainty throughout the perception process is needed to quantify the confidence level that the AI system has in its situational awareness and to improve accuracy.

Understanding theoretical capabilities and limitations of AI

While the ultimate goal for many AI algorithms is to address open challenges with human-like solutions, we do not have a good understanding of what the theoretical capabilities and limitations are for AI and the extent to which such human-like solutions are even possible with AI algorithms. Theoretical work is needed to better understand why AI techniques—especially machine learning—often work well in practice. While different disciplines (including mathematics, control sciences, and computer science) are studying this issue, the field currently lacks unified theoretical models or frameworks to understand AI system performance. Additional research is needed on computational solvability, which is an understanding of the classes of problems that AI algorithms are theoretically capable of solving, and likewise, those that they are not capable of solving. This understanding must be developed in the context of existing hardware, in order to see how the hardware affects the performance of these algorithms. Understanding which problems are theoretically unsolvable can lead researchers to develop approximate solutions to these problems, or even open up new lines of research on new hardware for AI systems. For example, when invented in the 1960s, Artificial Neural Networks (ANNs) could only be used to solve very simple problems. It only became feasible to use ANNs to solve complex problems after hardware improvements such as parallelization were made, and algorithms were adjusted to make use of the new hardware. Such developments were key factors in enabling today's significant advances in deep learning.

Pursuing research on general-purpose artificial intelligence

AI approaches can be divided into "narrow AI" and "general AI." Narrow AI systems perform individual tasks in specialized, well-defined domains, such as speech recognition, image recognition, and translation. Several recent, highly-visible, narrow AI systems, including IBM Watson and DeepMind's AlphaGo, have achieved major feats.[35,36] Indeed, these particular systems have been labeled "superhuman" because they have outperformed the best human players in Jeopardy! and Go, respectively. But these systems exemplify narrow AI, since they can only be applied to the tasks for which they are specifically designed. Using these systems on a wider range of problems requires a significant re-engineering effort. In contrast, the long-term goal of general AI is to create systems that

[35] In 2011, IBM Watson defeated two players considered among the best human players in the Jeopardy! game.

[36] In 2016, AlphaGo defeated the reigning world champion of Go, Lee Se-dol. Notably, AlphaGo combines deep learning and Monte Carlo search—a method developed in the 1980s—which itself builds on a probabilistic method discovered in the 1940s.

exhibit the flexibility and versatility of human intelligence in a broad range of cognitive domains, including learning, language, perception, reasoning, creativity, and planning. Broad learning capabilities would provide general AI systems the ability to transfer knowledge from one domain to another and to interactively learn from experience and from humans. General AI has been an ambition of researchers since the advent of AI, but current systems are still far from achieving this goal. The relationship between narrow and general AI is currently being explored; it is possible that lessons from one can be applied to improve the other and vice versa. While there is no general consensus, most AI researchers believe that general AI is still decades away, requiring a long-term, sustained research effort to achieve it.

Developing scalable AI systems

Groups and networks of AI systems may be coordinated or autonomously collaborate to perform tasks not possible with a single AI system, and may also include humans working alongside or leading the team. The development and use of such multi-AI systems creates significant research challenges in planning, coordination, control, and scalability of such systems. Planning techniques for multi-AI systems must be fast enough to operate and adapt in real time to changes in the environment. They should adapt in a fluid manner to changes in available communications bandwidth or system degradation and faults. Many prior efforts have focused on centralized planning and coordination techniques; however, these approaches are subject to single points of failure, such as the loss of the planner, or loss of the communications link to the planner. Distributed planning and control techniques are harder to achieve algorithmically, and are often less efficient and incomplete, but potentially offer greater robustness to single points of failure. Future research must discover more efficient, robust, and scalable techniques for planning, control, and collaboration of teams of multiple AI systems and humans.

Fostering research on human-like AI

Attaining human-like AI requires systems to explain themselves in ways that people can understand. This will result in a new generation of intelligent systems, such as intelligent tutoring systems and intelligent assistants that are effective in assisting people when performing their tasks. There is a significant gap, however, between the way current AI algorithms work and how people learn and perform tasks. People are capable of learning from just a few examples, or by receiving formal instruction and/or "hints" to performing tasks, or by observing other people performing those tasks. Medical schools take this approach, for example, when medical students learn by observing an established doctor performing a complex medical procedure. Even in high-performance tasks such as world-championship Go games, a master-level player would have played only a few thousand games to train him/herself. In contrast, it would take hundreds of years for a human to play the number of games needed to train AlphaGo. More foundational research on new approaches for achieving human-like AI would bring these systems closer to this goal.

Developing more capable and reliable robots

Significant advances in robotic technologies over the last decade are leading to potential impacts in a multiplicity of applications, including manufacturing, logistics, medicine, healthcare, defense and national security, agriculture, and consumer products. While robots were historically envisioned for static industrial environments, recent advances involve close collaborations between robots and humans. Robotics technologies are now showing promise in their ability to complement, augment, enhance, or emulate human physical capabilities or human intelligence. However, scientists need to make these robotic systems more capable, reliable, and easy-to-use.

Researchers need to better understand robotic perception to extract information from a variety of sensors to provide robots with real-time situational awareness. Progress is needed in cognition and reasoning to allow robots to better understand and interact with the physical world. An improved ability to adapt and learn will allow robots to generalize their skills, perform self-assessment of their current performance, and learn a repertoire of physical movements from human teachers. Mobility and manipulation are areas for further investigation so that robots can move across rugged and uncertain terrain and handle a variety of objects dexterously. Robots need to learn to team together in a seamless fashion and collaborate with humans in a way that is trustworthy and predictable.

Advancing hardware for improved AI

While AI research is most commonly associated with advances in software, the performance of AI systems has been heavily dependent on the hardware upon which it runs. The current renaissance in deep machine learning is directly tied to progress in GPU-based hardware technology and its improved memory,[37] input/output, clock speeds, parallelism, and energy efficiency. Developing hardware optimized for AI algorithms will enable even higher levels of performance than GPUs. One example is "neuromorphic" processors that are loosely inspired by the organization of the brain and, in some cases, optimized for the operation of neural networks.[38]

Hardware advances can also improve the performance of AI methods that are highly data-intensive. Further study of methods to turn on and off data pipelines in controlled ways throughout a distributed system is called for. Continued research is also needed to allow machine learning algorithms to efficiently learn from high-velocity data, including distributed machine learning algorithms that simultaneously learn from multiple data pipelines. More advanced machine learning-based feedback methods will allow AI systems to intelligently sample or prioritize data from large-scale simulations, experimental instruments, and distributed sensor systems, such as Smart Buildings and the Internet of Things (IoT). Such methods may require dynamic I/O decision-making, in which choices are made in real time to store data based on importance or significance, rather than simply storing data at fixed frequencies.

Creating AI for improved hardware

While improved hardware can lead to more capable AI systems, AI systems can also improve the performance of hardware.[39] This reciprocity will lead to further advances in hardware performance, since physical limits on computing require novel approaches to hardware designs.[40] AI-based methods could be especially important for improving the operation of high-performance computing (HPC) systems. Such systems consume vast quantities of energy. AI is being used to predict HPC performance and resource usage, and to make online optimization decisions that increase efficiency; more advanced AI techniques could further enhance system performance. AI can also be used to create

[37] GPU stands for graphics processing unit, which is a power- and cost-efficient processor incorporating hundreds of processing cores; this design makes it especially well suited for inherently parallel applications, including most AI systems.

[38] Neuromorphic computing refers to the ability of hardware to learn, adapt, and physically reconfigure, taking inspiration from biology or neuroscience.

[39] M. Milano and L. Benini, "Predictive Modeling for Job Power Consumption in HPC Systems," In *Proceedings of High Performance Computing: 31st International Conference, ISC High Performance 2016* (Springer Vol. 9697, 2016).

[40] These physical limits on computing are called *Dennard scaling*, and lead to high on-chip power densities and the phenomenon called "dark silicon", where different parts of a chip need to be turned off in order to limit temperatures and ensure data integrity.

self-reconfigurable HPC systems that can handle system faults when they occur, without human intervention.[41]

Improved AI algorithms can increase the performance of multi-core systems by reducing data movements between processors and memory—the primary impediment to exascale computing systems that operate 10 times faster than today's supercomputers.[42] In practice, the configuration of executions in HPC systems are never the same, and different applications are executed concurrently, with the state of each different software code evolving independently in time. AI algorithms need to be designed to operate online and at scale for HPC systems.

[41] A. Cocaña-Fernández, J. Ranilla, and L. Sánchez, "Energy-efficient allocation of computing node slots in HPC clusters through parameter learning and hybrid genetic fuzzy system modeling," *Journal of Supercomputing* 71 (2015):1163-1174.

[42] Exascale computing systems can achieve at least a billion billion calculations per second.

Strategy 2: Develop Effective Methods for Human-AI Collaboration

2019 Update	**Developing AI systems that complement and augment human capabilities, with increasing focus on the future of work**

Since the release of the 2016 *National AI R&D Strategic Plan*, national interest has grown in human-AI collaboration. When AI systems complement and augment human capabilities, humans and AI become partners across a range of shared to fully autonomous scenarios. In particular, human-AI collaboration has been elevated as both a challenge and an opportunity in the context of the future of work.

In the past three years, newly established as well as longstanding conferences, workshops, and task forces have prioritized human-AI collaboration broadly. For example, the Conference on Human Computation and Crowdsourcing has grown from a workshop to a major international conference that fosters research in the intersection of AI and human-computer interaction (HCI).[45] In 2018, the Association for the Advancement of Artificial Intelligence selected human-AI collaboration as the emerging topic for its annual conference.[46] In May 2019, the largest conference on human-computer interaction, CHI, included a workshop on "Bridging the Gap Between AI and HCI."[47] The journal *Human-Computer Interaction* put out a call in March 2019 for submissions for a special issue on "unifying human-computer interaction and artificial intelligence."[48]

Human-AI Collaboration: Recent agency R&D programs

Since the release of the 2016 *National AI R&D Strategic Plan*, several agencies have initiated efforts for Strategy 2:

- NSF's Future of Work at the Human-Technology Frontier[43] Big Idea is supporting socio-technical research enabling a future where intelligent technologies collaborate synergistically with humans to achieve broad participation in the workforce and improve the social, economic, and environmental benefits across a range of work settings.
- NOAA (National Oceanographic and Atmospheric Administration) is advancing human-AI collaboration for hurricane, tornado, and other severe weather predictions where the human forecaster and an AI system work together to improve severe weather warning generation and to identify distinct patterns that are precursors to extreme events. Sometimes referred to as "humans above the loop," human forecasters oversee the AI system's predictions and direct the outcomes.
- NIH has ongoing research in natural language processing based on a database of 96.3 million facts extracted from all MEDLINE citations maintained by the National Library of Medicine.
- A 2019 DOE workshop report on Scientific Machine Learning identified priority research directions, major scientific use cases, and the emerging trend that human-AI collaborations will transform the way science is done.[44]

[43] https://www.nsf.gov/eng/futureofwork.jsp

[44] DOE workshop report, *Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*: https://www.osti.gov/biblio/1478744.

[45] Welcome to HCOMP 2019: https://www.humancomputation.com/.

[46] AAAI-18 Emerging Topic Human-AI Collaboration: http://www.aaai.org/Conferences/AAAI/2018/aaai18emergingcall.php.

[47] Where is the Human? Bridging the Gap Between AI and HCI: CHI 2019 Workshop: https://michae.lv/ai-hci-workshop/.

[48] Call: "Unifying Human Computer Interaction and Artificial Intelligence" issue of *Human-Computer Interaction*: https://ispr.info/2019/02/20/call-unifying-human-computer-interaction-and-artificial-intelligence-issue-of-human-computer-interaction/.

In the context of work, conferences have emerged exploring the role of the human, the machine, and their partnership, such as MIT's Computer Science and Artificial Intelligence Lab (CSAIL) and the Initiative on the Digital Economy that launched the Annual AI and the Future of Work Congress.[49,50] As part of *A 20-Year Community Roadmap for Artificial Intelligence Research in the U.S.*,[51] in 2019 the Computing Community Consortium (CCC) held a workshop focused on meaningful interaction between humans and AI systems.[52] Additionally, the CCC operated the Human Technology Frontier task force in 2017-2018 to focus on the potential of technology to augment human performance in, including but not limited to, the workplace, the classroom, and the healthcare system.[53]

The cross-strategy principle in the 2016 *National AI R&D Strategic Plan*, "appropriate trust of AI systems requires explainability, especially as the AI grows in scale and complexity," has seen an R&D call to action in the context of human-AI collaborations. This principle has been identified by a number of professional societies and agencies as a priority area (see sidebar). This research area reflects the intersection of Strategies 2 and 3, as explainability, fairness, and transparency are key principles for AI systems to effectively collaborate with humans. Likewise, the challenge of understanding and designing human-AI ethics and value alignment into systems remains an open research area. In parallel, the private sector has responded with principles for effective human-AI collaboration.[54,55]

As Federal agencies have increased AI investments in the past three years along mission objectives, they have shared a common emphasis on human-machine cognition, autonomy, and agency, such as in decision support, risk modeling, situational awareness, and trusted machine intelligence (see sidebar). Through such R&D investments, research partnerships are growing across a number of axes, bringing together computational scientists; behavioral, cognitive, and psychological scientists; and scientists and engineers from other domains. New collaborations have formed between academic researchers and users of AI systems inside and outside the workplace.

Moving forward, it is critical that Federal agencies continue to foster AI R&D in the open world to promote the design of AI systems that incorporate and accommodate the situations and goals of users so that AI systems and users can work collaboratively in both anticipated and unanticipated circumstances.

While completely autonomous AI systems will be important in some application domains (e.g., underwater or deep space exploration), many other application areas (e.g., disaster recovery and medical diagnostics) are most effectively addressed by a combination of humans and AI systems working together to achieve application goals. This collaborative interaction takes advantage of the complementary nature of humans and AI systems. While effective approaches for human-AI collaboration already exist, most of these are "point solutions" that only work in specific environments using specific platforms toward specific goals. Generating point solutions for every possible application instance does not scale; more work is thus needed to go beyond these point solutions toward more

[49] https://futureofwork.csail.mit.edu/.

[50] AI and Future of Work Innovation Summit 2019: https://analyticsevent.com/.

[51] https://cra.org/ccc/wp-content/uploads/sites/2/2019/03/AI_Roadmap_Exec_Summary-FINAL-.pdf

[52] Artificial Intelligence Roadmap Workshop 2 – Interaction: https://cra.org/ccc/events/artificial-intelligence-roadmap-workshop-2-interaction/.

[53] https://cra.org/ccc/human-technology-frontier/

[54] https://www.microsoft.com/en-us/research/uploads/prod/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf

[55] https://www.partnershiponai.org/about/#our-work

general methods of human-AI collaboration. The tradeoffs must be explored between designing general systems that work in all types of problems, requiring less human effort to build and greater facility for switching between applications, versus building a large number of problem-specific systems that may work more effectively for each problem.

Future applications will vary considerably in the functional role divisions between humans and AI systems, the nature of the interactions between humans and AI systems, the number of humans and other AI systems working together, and how humans and AI systems will communicate and share situational awareness. Functional role divisions between humans and AI systems typically fall into one of the following categories:

1. *AI performs functions alongside the human*: AI systems perform peripheral tasks that support the human decision maker. For example, AI can assist humans with working memory, short or long-term memory retrieval, and prediction tasks.
2. *AI performs functions when the human encounters high cognitive overload*: AI systems perform complex monitoring functions (such as ground proximity warning systems in aircraft), decision making, and automated medical diagnoses when humans need assistance.
3. *AI performs functions in lieu of a human*: AI systems perform tasks for which humans have very limited capabilities, such as for complex mathematical operations, control guidance for dynamic systems in contested operational environments, aspects of control for automated systems in harmful or toxic environments, and in situations where a system should respond very rapidly (e.g., in nuclear reactor control rooms).

Achieving effective interactions between humans and AI systems requires additional R&D to ensure that the system design does not lead to excessive complexity, undertrust, or overtrust. The familiarity of humans with AI systems can be increased through training and experience, to ensure that the human has a good understanding of the AI system's capabilities and what the AI system can and cannot do. To address these concerns, certain human-centered automation principles should be used in the design and development of these systems:[56]

1. Employ intuitive, user-friendly design of human-AI system interfaces, controls, and displays.
2. Keep the operator informed. Display critical information, states of the AI system, and changes to these states.
3. Keep the operator trained. Engage in recurrent training for general knowledge, skills, and abilities (KSAs), as well as training in algorithms and logic employed by AI systems and the expected failure modes of the system.
4. Make automation flexible. Deploying AI systems should be considered as a design option for operators who wish to decide whether they want to use them or not. Also important is the design and deployment of adaptive AI systems that can be used to support human operators during periods of excessive workload or fatigue.[57,58]

Many fundamental challenges arise for researchers when creating systems that work effectively with humans. Several of these important challenges are outlined in the following subsections.

[56] C. Wickens and J. G. Hollands, "Attention, time-sharing, and workload." In *Engineering, Psychology and Human Performance* (London: Pearson PLC, 1999), 439-479.

[57] https://www.nasa.gov/mission_pages/SOFIA/index.html

[58] https://cloud1.arc.nasa.gov/intex-na/

Seeking new algorithms for human-aware AI

Over the years, AI algorithms have become able to solve problems of increasing complexity. However, there is a gap between the capabilities of these algorithms and the usability of these systems by humans. *Human-aware* intelligent systems are needed that can interact intuitively with users and enable seamless machine-human collaborations. Intuitive interactions include shallow interactions, such as when a user discards an option recommended by the system; model-based approaches that take into account the users' past actions; or even deep models of user intent that are based upon accurate human cognitive models. Interruption models must be developed that allow an intelligent system to interrupt the human only when necessary and appropriate. Intelligent systems should also have the ability to augment human cognition, knowing which information to retrieve when the user needs it, even when they have not prompted the system explicitly for that information. Future intelligent systems must be able to account for human social norms and act accordingly. Intelligent systems can more effectively work with humans if they possess some degree of emotional intelligence, so that they can recognize their users' emotions and respond appropriately. An additional research goal is to go beyond interactions of one human and one machine, toward a "systems-of-systems", that is, teams composed of multiple machines interacting with multiple humans.

Human-AI system interactions have a wide range of objectives. AI systems need the ability to represent a multitude of goals, actions that they can take to reach those goals, constraints on those actions, and other factors, as well as easily adapt to modifications in the goals. In addition, humans and AI systems must share common goals and have a mutual understanding of them and relevant aspects of their current states. Further investigation is needed to generalize these facets of human-AI systems to develop systems that require less human engineering.

Developing AI techniques for human augmentation

While much of the prior focus of AI research has been on algorithms that match or outperform people performing narrow tasks, more work is needed to develop systems that augment human capabilities across many domains. Human augmentation research includes algorithms that work on a stationary device (such as a computer); wearable devices (such as smart glasses); implanted devices (such as brain interfaces); and in specific user environments (such as specially tailored operating rooms). For example, augmented human awareness could enable a medical assistant to point out a mistake in a medical procedure, based on data readings combined from multiple devices. Other systems could augment human cognition by helping the user recall past experiences applicable to the user's current situation.

Another type of collaboration between humans and AI systems involves active learning for intelligent data understanding. In active learning, input is sought from a domain expert and learning is only performed on data when the learning algorithm is uncertain. This is an important technique to reduce the amount of training data that needs to be generated in the first place, or the amount that needs to be learned. Active learning is also a key way to obtain domain expert input and increase trust in the learning algorithm. Active learning has so far only been used within supervised learning; further research is needed to incorporate active learning into unsupervised learning (e.g., clustering, anomaly detection) and reinforcement learning.[59] Probabilistic networks allow domain knowledge to be included in the form of prior probability distributions. General ways of allowing machine learning algorithms to incorporate domain knowledge must be sought, whether in the form of mathematical models, text, or others.

[59] While supervised learning requires humans to provide the ground-truth answers, reinforcement learning and unsupervised learning do not.

Developing techniques for visualization and human-AI interfaces

Better visualization and user interfaces are additional areas that need much greater development to help humans understand large-volume modern datasets and information coming from a variety of sources. Visualization and user interfaces must clearly present increasingly complex data and information derived from them in a human-understandable way. Providing real-time results is important in safety-critical operations and may be achieved with increasing computational power and connected systems. In these types of situations, users need visualization and user interfaces that can quickly convey the correct information for real-time response.

Human-AI collaboration can be applied in a wide variety of environments, and where there are constraints on communication. In some domains, human-AI communication latencies are low and communication is rapid and reliable. In other domains (e.g., NASA's deployment of the rovers Spirit and Opportunity to Mars), remote communication between humans and the AI system has a very high latency (e.g., round trip times of 5-20 minutes between Earth and Mars), thus requiring the deployed platform(s) to operate largely autonomously, with only high-level strategic goals communicated to the platform. These communications requirements and constraints are important considerations for the R&D of user interfaces.

Developing more effective language processing systems

Enabling people to interact with AI systems through spoken and written language has long been a goal of AI researchers. While significant advances have been made, considerable open research challenges must be addressed in language processing before humans can communicate as effectively with AI systems as they do with other humans. Much recent progress in language processing has been credited to the use of data-driven machine learning approaches, which have resulted in successful systems that, for example, successfully recognize fluent English speech in quiet surroundings in real time. These achievements, however, are only first steps toward reaching longer-term goals. Current systems cannot deal with real-world challenges such as speech in noisy surroundings, heavily accented speech, children's speech, impaired speech, and speech for sign languages. The development of language processing systems capable of engaging in real-time dialogue with humans is also needed. Such systems will need to infer the goals and intentions of its human interlocutors, use the appropriate register, style and rhetoric for the situation, and employ repair strategies in case of dialogue misunderstandings. Further research is needed on developing systems that more easily generalize across different languages. Additionally, more study is required on acquiring useful structured domain knowledge in a form readily accessible by language processing systems.

Language processing advances in many other areas are also needed to make interactions between humans and AI systems more natural and intuitive. Robust computational models must be built for patterns in both spoken and written language that provide evidence for emotional state, affect, and stance, and for determining the information that is implicit in speech and text. New language processing techniques are needed for grounding language in the environmental context for AI systems that operate in the physical world, such as in robotics. Finally, since the manner in which people communicate in online interactions can be quite different from voice interactions, models of languages used in these contexts must be perfected so that social AI systems can interact more effectively with people.

Strategy 3: Understand and Address the Ethical, Legal, and Societal Implications of AI

2019 Update	**Addressing ethical, legal, and societal considerations in AI**

Since the release of the 2016 *National AI R&D Strategic Plan*, R&D activities addressing the ethical, legal, and societal implications of AI system development and deployment have increased. There is a growing realization that AI systems must be "trustworthy," and that AI could transform many sectors of social and economic life, including employment, healthcare, and manufacturing. International organizations such as the Organisation for Economic Co-operation and Development (OECD)[63] and the G7 Innovation Ministers[64] have encouraged R&D to increase trust in and adoption of AI.

The 2016 *National AI R&D Strategic Plan* was prescient in identifying research themes in privacy; improving fairness, transparency, and accountability of AI systems by design; and designing architectures for ethical AI. Research conferences dedicated to fairness, accountability, and transparency in ML and AI systems have flourished.[65] Federal agencies have responded with a variety of new research programs and meetings focused on these critical areas (see sidebar).

Explainability, fairness, and transparency: Recent agency R&D programs

Since the release of the 2016 *National AI R&D Strategic Plan*, a number of agencies have initiated AI R&D programs for Strategy 3:

- DARPA's Explainable AI (XAI) program[60] aims to create a suite of ML techniques that produce more explainable AI systems while maintaining a high level of learning performance (prediction accuracy). XAI will also enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems. More generally, DoD is committed to "leading in military ethics and AI safety" as one of five key actions outlined in the strategic approach that guides its efforts to accelerate the adoption of AI systems.[61]
- NSF and Amazon are collaborating[62] to jointly support research focused on AI fairness with the goal of contributing to trustworthy AI systems that are readily accepted and deployed to tackle grand challenges facing society. Specific topics of interest include, but are not limited to, transparency, explainability, accountability, potential adverse biases and effects, mitigation strategies, validation of fairness, and considerations of inclusivity.

[60] https://www.darpa.mil/program/explainable-artificial-intelligence

[61] "Summary of the 2018 Department of Defense Artificial Intelligence Strategy": https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.

[62] https://www.nsf.gov/pubs/2019/nsf19571/nsf19571.htm

[63] "OECD Initiatives on AI": http://www.oecd.org/going-digital/ai/oecd-initiatives-on-ai.htm.

[64] "G7 Innovation Ministers' Statement on AI": http://www.g8.utoronto.ca/employment/2018-labour-annex-b-en.html.

[65] http://www.fatml.org/; https://fatconference.org/; http://www.aies-conference.com/

The 2019 *Executive Order on Maintaining American Leadership in Artificial Intelligence* emphasizes that maintaining American leadership in AI requires a concerted effort to promote advancements in technology and innovation, while protecting civil liberties, privacy, and American values:[1]

> The United States must foster public trust and confidence in AI technologies and protect civil liberties, privacy, and American values in their application in order to fully realize the potential of AI technologies for the American people.

More R&D is needed to develop AI architectures that incorporate ethical, legal, and societal concerns through technical mechanisms such as transparency and explainability. This R&D will require intensive collaboration among technical experts as well as stakeholders and specialists in other fields including the social and behavioral sciences, law, ethics, and philosophy. Since ethical decisions may also be heavily context- or application-dependent, collaboration with domain experts could be required as well. This interdisciplinary approach could be incorporated in the training, design, testing, evaluation, and implementation of AI in the interests of understanding and accounting for AI-induced decisions and actions and mitigating unintended consequences.

Federal agencies should therefore continue to foster the growing community of interest in further R&D of these issues by sponsoring research and convening experts and stakeholders.

When AI agents act autonomously, we expect them to behave according to the formal and informal norms to which we hold our fellow humans. As fundamental social ordering forces, law and ethics therefore both inform and adjudge the behavior of AI systems. The dominant research needs involve both understanding the ethical, legal, and social implications of AI, as well as developing methods for AI design that align with ethical, legal, and social principles. Privacy concerns must also be taken into account; further information on this issue can be found in the *National Privacy Research Strategy*.[66]

As with any technology, the acceptable uses of AI will be informed by the tenets of law and ethics; the challenge is how to apply those tenets to this new technology, particularly those involving autonomy, agency, and control.

As illuminated in "Research Priorities for Robust and Beneficial Artificial Intelligence,"[67]

> In order to build systems that robustly behave well, we of course need to decide what good behavior means in each application domain. This ethical dimension is tied intimately to questions of what engineering techniques are available, how reliable these techniques are, and what trade-offs are made—all areas where computer science, machine learning, and broader AI expertise is valuable.

Research in this area can benefit from multidisciplinary perspectives that involve experts from computer science, social and behavioral sciences, ethics, biomedical science, psychology, economics, law, and policy research. Further investigation is needed in areas both inside and outside of the NITRD-relevant IT domain (i.e., in information technology, as well as in the disciplines mentioned previously) to inform the R&D and use of AI systems and their impacts on society.

The following subsections explore key information technology research challenges in this area.

---

[66] https://www.nitrd.gov/pubs/NationalPrivacyResearchStrategy.pdf

[67] "An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence" (Future of Life Institute): http://futureoflife.org/ai-open-letter/.

Improving fairness, transparency, and accountability by design

Many concerns have been voiced about the susceptibility of data-intensive AI algorithms to error and misuse, and the possible ramifications for gender, age, racial, or economic classes. The proper collection and use of data for AI systems, in this regard, represent an important challenge. Beyond purely data-related issues, however, larger questions arise about the design of AI to be inherently just, fair, transparent, and accountable. Researchers must learn how to design these systems so that their actions and decision-making are transparent and easily interpretable by humans, and thus can be examined for any bias they may contain, rather than just learning and repeating these biases. There are serious intellectual issues about how to represent and "encode" value and belief systems. Scientists must also study to what extent justice and fairness considerations can be designed into the system, and how to accomplish this within the bounds of current engineering techniques.

Building ethical AI

Beyond fundamental assumptions of justice and fairness are other concerns about whether AI systems can exhibit behavior that abides by general ethical principles. How might advances in AI frame new "machine-relevant" questions in ethics, or what uses of AI might be considered unethical? Ethics is inherently a philosophical question while AI technology depends on, and is limited by, engineering. Within the limits of what is technologically feasible, therefore, researchers must strive to develop algorithms and architectures that are verifiably consistent with, or conform to, existing laws, social norms and ethics—clearly a very challenging task. Ethical principles are typically stated with varying degrees of vagueness and are hard to translate into precise system and algorithm design. There are also complications when AI systems, particularly with new kinds of autonomous decision-making algorithms, face moral dilemmas based on independent and possibly conflicting value systems. Ethical issues vary according to culture, religion, and beliefs. However, acceptable ethics reference frameworks can be developed to guide AI system reasoning and decision-making in order to explain and justify its conclusions and actions. A multidisciplinary approach is needed to generate datasets for training that reflect an appropriate value system, including examples that indicate preferred behavior when presented with difficult moral issues or with conflicting values. These examples can include legal or ethical "corner cases," labeled by an outcome or judgment that is transparent to the user.[68] AI needs adequate methods for values-based conflict resolution, where the system incorporates principles that can address the realities of complex situations where strict rules are impracticable.

Designing architectures for ethical AI

Additional progress in fundamental research must be made to determine how to best design architectures for AI systems that incorporate ethical reasoning. A variety of approaches have been suggested, such as a two-tier monitor architecture that separates the operational AI from a monitor agent that is responsible for the ethical or legal assessment of any operational action.[68] An alternative view is that safety engineering is preferred, in which a precise conceptual framework for the AI agent architecture is used to ensure that AI behavior is safe and not harmful to humans.[69] A third method is to formulate an ethical architecture using set theoretic principles, combined with logical constraints

[68] A. Etzioni and O. Etzioni, "Designing AI Systems that Obey Our Laws and Values," *Communications of the ACM* 59(9) (2016):29-31.

[69] R. Y. Yampolsky, "Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach." In *Philosophy and Theory of Artificial Intelligence,* ed. V.C. Muller (Heidelberg: Springer Verlag, 2013), 389-396.

on AI system behavior that restrict action to conform to ethical doctrine.[70] As AI systems become more general, their architectures will likely include subsystems that can take on ethical issues at multiple levels of judgment, including:[71] rapid response pattern matching rules, deliberative reasoning for slower responses for describing and justifying actions, social signaling to indicate trustworthiness for the user, and social processes that operate over even longer time scales to enable the system to abide by cultural norms. Researchers will need to focus on how to best address the overall design of AI systems that align with ethical, legal, and societal goals.

[70] R. C. Arkin, "Governing Legal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture," Georgia Institute of Technology Technical Report, GIT-GVU-07-11, 2007.

[71] B. Kuipers, "Human-like Morality and Ethics for Robots," AAAI-16 Workshop on AI, Ethics and Society, 2016; https://web.eecs.umich.edu/~kuipers/papers/Kuipers-aaaiws-16.pdf

Strategy 4: Ensure the Safety and Security of AI Systems

2019 Update	**Creating robust and trustworthy AI systems**

Since the 2016 release of the *National AI R&D Strategic Plan*, there has been rapid growth in scientific and societal understanding of AI security and safety. Much of this new knowledge has helped identify new problems: it is more evident now how AI systems can be made to do the wrong thing, learn the wrong thing, or reveal the wrong thing, for example, through adversarial examples, data poisoning, and model inversion, respectively. Unfortunately, technical solutions for these AI safety and security problems remain elusive.

To address all of these problems, the safety and security of AI systems must be considered in all stages of the AI system lifecycle, from the initial design and data/model building, to verification and validation, deployment, operation, and monitoring. Indeed, the notion of "safety (or security) by design" might impart an incorrect notion that these are only concerns of system designers; instead, they must be considered throughout the system lifecycle, not just at the design stage, and so must be an important part of the AI R&D portfolio.

When AI components are connected to other systems or information that must be safe or secure, the AI vulnerabilities and performance requirements (e.g., very low false-positive and false-negative rates, when operating over high volumes of data)

AI safety and security: Recent agency R&D programs

Since the release of the 2016 *National AI R&D Strategic Plan*, a number of agencies have initiated efforts supporting Strategy 4:

- DOT published new Federal guidance for automated vehicles in October 2018 supporting the safe integration of automation into the broad multimodal surface transportation system. *Preparing for the Future of Transportation: Automated Vehicles 3.0*[72] advances DOT's principles for safe integration of automated vehicles. The document also reiterates prior safety guidance, provides new multimodal safety guidance, and outlines a process for working with DOT as this new technology evolves. As of May 2019, fourteen companies had released Voluntary Safety Self-Assessments detailing how they will incorporate safety into their design and testing of automated driving systems.[73]
- In December 2018, IARPA announced two programs on AI security: Secure, Assured, Intelligent Learning Systems (SAILS)[74] and Trojans in Artificial Intelligence (TrojAI).[75] DARPA announced another program in February 2019, Guaranteeing AI Robustness against Deception (GARD).[76] Together, these programs aim to combat a range of attacks on AI systems.
- As noted in Strategy 3, DoD is committed to "leading in military ethics and AI safety" as one of five key actions outlined in the strategic approach that guides its efforts to accelerate the adoption of AI systems.[77]

[72] https://www.transportation.gov/av/3

[73] https://www.nhtsa.gov/automated-driving-systems/voluntary-safety-self-assessment

[74] https://www.iarpa.gov/index.php/research-programs/sails

[75] https://www.iarpa.gov/index.php/research-programs/trojai

[76] https://www.darpa.mil/news-events/2019-02-06

[77] "Summary of the 2018 Department of Defense Artificial Intelligence Strategy": https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.

are inherited by the larger systems. These challenges are not static; as AI systems continue to grow in capabilities, they will likely grow in complexity, making it ever harder for correct performance or privacy of information to be verified and validated. This complexity may also make it increasingly difficult to explain decisions in ways that justify high levels of trust from human users (see Strategy 3).

Making AI trustworthy—now and into the future—is a critical issue that requires Federal Government R&D investments (see sidebar), along with collaborative efforts among government, industry, academia, and civil society. Engineering trustworthy AI systems may benefit from borrowing existing practices in safety engineering in other fields that have learned how to account for potential misbehavior of non-AI autonomous or semi-autonomous systems. However, AI-specific problems mean that novel techniques for program analysis, testing, formal verification, and synthesis will be critical to establish that an AI-based system meets its specifications—that is, that the system does exactly what it is supposed to do and no more. These problems are exacerbated in AI-based systems that can be easily fooled, evaded, and misled in ways that can have profound security implications. An emerging research area is adversarial ML, which explores both the analysis of vulnerabilities in ML algorithms as well as algorithmic techniques that yield more robust learning. Well-known attacks on ML include adversarial classifier evasion attacks, where the attacker changes behavior to escape being detected, and poisoning attacks, where training data itself is corrupted. There is growing need for research that systematically explores the space of adversaries that attack ML and other AI-based systems and to design algorithms that provide provable robustness guarantees against classes of adversaries.

Methods must be developed to make safe and secure the creation, evaluation, deployment, and containment of AI, and these methods must scale to match the capability and complexity of AI. Evaluating these methods will require new metrics, control frameworks, and benchmarks for testing and assessing the safety of increasingly powerful systems. Both methods and metrics must incorporate human factors, with safe AI objectives defined by human designers' goals, safe AI operations defined by human users' habits, and safe AI metrics defined by human evaluators' understanding. Producing human-driven and human-understandable methods and metrics for the safety of AI systems will enable policymakers, the private sector, and the public to accurately judge the evolving AI safety landscape and appropriately proceed within it.

Before an AI system is put into widespread use, assurance is needed that the system will operate safely and securely, in a controlled manner. Research is needed to address this challenge of creating AI systems that are reliable, dependable, and trustworthy. As with other complex systems, AI systems face important safety and security challenges due to:[78]

- *Complex and uncertain environments*: In many cases, AI systems are designed to operate in complex environments, with a large number of potential states that cannot be exhaustively examined or tested. A system may confront conditions that were never considered during its design.
- *Emergent behavior*: For AI systems that learn after deployment, a system's behavior may be determined largely by periods of learning under unsupervised conditions. Under such conditions, it may be difficult to predict a system's behavior.
- *Goal misspecification*: Due to the difficulty of translating human goals into computer instructions, the goals that are programmed for an AI system may not match the goals that were intended by the programmer.

[78] J. Bornstein, "DoD Autonomy Roadmap – Autonomy Community of Interest," Presentation at NDIA 16th Annual Science & Engineering Technology Conference, March 2015.

- *Human-machine interactions*: In many cases, the performance of an AI system is substantially affected by human interactions. In these cases, variation in human responses may affect the safety of the system.[79]

To address these issues and others, additional investments are needed to advance AI safety and security,[80] including explainability and transparency, trust, verification and validation, security against attacks, and long-term AI safety and value-alignment.

Improving explainability and transparency

A key research challenge is increasing the "explainability" or "transparency" of AI. Many algorithms, including those based on deep learning, are opaque to users, with few existing mechanisms for explaining their results. This is especially problematic for domains such as healthcare, where doctors need explanations to justify a particular diagnosis or a course of treatment. AI techniques such as decision-tree induction provide built-in explanations but are generally less accurate. Thus, researchers must develop systems that are transparent, and intrinsically capable of explaining the reasons for their results to users.

Building trust

To achieve trust, AI system designers need to create accurate, reliable systems with informative, user-friendly interfaces, while the operators must take the time for adequate training to understand system operation and limits of performance. Complex systems that are widely trusted by users, such as manual controls for vehicles, tend to be transparent (the system operates in a manner that is visible to the user), credible (the system's outputs are accepted by the user), auditable (the system can be evaluated), reliable (the system acts as the user intended), and recoverable (the user can recover control when desired). A significant challenge to current and future AI systems remains the inconsistent quality of software production technology. As advances bring greater linkages between humans and AI systems, the challenge in the area of trust is to keep pace with changing and increasing capabilities, anticipate technological advances in adoption and long-term use, and establish governing principles and policies for the study of best practices for design, construction, and use, including proper operator training for safe operation.

Enhancing verification and validation

New methods are needed for verification and validation of AI systems. "Verification" establishes that a system meets formal specifications, while "validation" establishes that a system meets the user's operational needs. Safe AI systems may require new means of *assessment* (determining if the system is malfunctioning, perhaps when operating outside expected parameters), *diagnosis* (determining the causes for the malfunction), and *repair* (adjusting the system to address the malfunction). For systems operating autonomously over extended periods of time, system designers may not have considered every condition the system will encounter. Such systems may need to possess capabilities for self-assessment, self-diagnosis, and self-repair in order to be robust and reliable.

---

[79] J. M. Bradshaw, R. R. Hoffman, M. Johnson, and D. D. Woods, "The Seven Deadly Myths of Autonomous Systems," *IEEE Intelligent Systems* 28(3)(2013):54-61.

[80] See, for instance: D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mane, "Concrete Problems in AI Safety," 2016, arXiv: 1606.06565v2; S. Russell, D. Dewey, and M. Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence," 2016, arXiv: 1602.03506; T. G. Dietterich and E. J. Horvitz, "Rise of Concerns about AI: Reflections and Directions," *Communications of the ACM*, 58(10)(2015); and K. Sotala and R. Yampolskiy, "Responses to catastrophic AGI risk: A survey," *Physica Scripta*, 90(1), 19 December 2014.

Securing against attacks

AI embedded in critical systems must be robust in order to handle accidents but should also be secure to a wide range of intentional cyber attacks. Security engineering involves understanding the vulnerabilities of a system and the actions of actors who may be interested in attacking it. While cybersecurity R&D needs are addressed in greater detail in the NITRD 2016 *Federal Cybersecurity R&D Strategic Plan*,[81] some cybersecurity risks are specific to AI systems. For example, one key research area is "adversarial machine learning" that explores the degree to which AI systems can be compromised by "contaminating" training data, by modifying algorithms, or by making subtle changes to an object that prevent it from being correctly identified (e.g., prosthetics that spoof facial recognition systems). The implementation of AI in cybersecurity systems that require a high degree of autonomy is also an area for further study. One recent example of work in this area is DARPA's Cyber Grand Challenge that involved AI agents autonomously analyzing and countering cyber attacks.[82]

Achieving long-term AI safety and value-alignment

AI systems may eventually become capable of "recursive self-improvement," in which substantial software modifications are made by the software itself, rather than by human programmers. To ensure the safety of self-modifying systems, additional research is called for to develop: self-monitoring architectures that check systems for behavioral consistency with the original goals of human designers; confinement strategies for preventing the release of systems while they are being evaluated; value learning, in which the values, goals, or intentions of users can be inferred by a system; and value frameworks that are provably resistant to self-modification.

[81] https://www.nitrd.gov/pubs/2016-Federal-Cybersecurity-Research-and-Development-Strategic-Plan.pdf; this is being updated in 2019.

[82] https://archive.darpa.mil/CyberGrandChallenge_CompetitorSite/

Strategy 5: Develop Shared Public Datasets and Environments for AI Training and Testing

2019 Update	**Increasing access to datasets and associated challenges**

At the time of the 2016 *National AI R&D Strategic Plan*'s release, publicly available datasets and environments were already playing a critical role in pushing forward AI R&D, particularly in areas such as computer vision, natural language processing, and speech recognition. ImageNet,[84] with more than 14 million labeled objects, along with associated computer vision community challenges (e.g., the ImageNet Large Scale Visual Recognition Challenge[85] that evaluates algorithms for object detection and image classification), have played an especially vital role in the community. As translational applications for ML are being found in myriad application areas such as healthcare, medicine, and smart and connected communities, the need has grown for publicly available datasets in domain-specific areas.

The importance of datasets and models – in particular, those of the Federal Government – is explicitly called out in the 2019 *Executive Order on Maintaining American Leadership in Artificial Intelligence*:[1]

> Heads of all agencies shall review their Federal data and models to identify opportunities to increase access and use by the greater non-Federal AI research community in a manner that benefits that community, while protecting safety, security, privacy, and confidentiality. Specifically, agencies shall improve data and model inventory documentation to enable discovery and usability, and shall

Shared Public Datasets and Environments for AI Training and Testing: Recent agency R&D programs

Since the release of the 2016 *National AI R&D Strategic Plan*, a number of agencies have initiated efforts supporting Strategy 5:

- DOT sponsored the Second Strategic Highway Research Program (SHRP2) Naturalistic Driving Study (NDS),[83] which recorded more than 5.4 million trips taken by more than 3,400 drivers and vehicles. An in-vehicle data acquisition system (DAS) unit gathered and stored data from forward radar, four video cameras, accelerometers, vehicle network information, a geographic positioning system, and an onboard lane tracker. Data from the DAS were recorded continuously while participants' vehicles were operating. Whereas summaries of the NDS data are public, access to the detailed datasets requires qualified research ethics training.
- The VA Data Commons is creating the largest linked medical–genomics dataset in the world with tools for enabling ML and AI, and guided by veterans' preferences. This effort is leveraging applicable NIST standards, laws, and executive orders.
- GSA (General Services Administration) is working to enable the use of cloud computing resources for federally funded AI R&D. Data.gov and code.gov, housed at GSA, contain over 246,000 datasets and code from across agencies and automatically harvest datasets released by agencies.
- The NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) initiative, a partnership with industry-leading cloud service providers, is enabling researcher access to major data assets that are funded across NIH and that are stored in cloud environments.

[83] https://insight.shrp2nds.us/
[84] http://www.image-net.org/
[85] http://www.image-net.org/challenges/LSVRC/

prioritize improvements to access and quality of AI data and models based on the AI research community's feedback.

A new NSTC Subcommittee on Open Science was created in 2018 to coordinate Federal efforts on open and FAIR (findable, accessible, interoperable, and reusable) data. R&D investments will be needed to develop tools and resources that make it easier to identify, use, and manipulate relevant datasets (including Federal datasets), verify data provenance, and respect appropriate use policy. Many of these datasets themselves may be of limited use in an AI context without an investment in labeling and curation. Federal agencies should engage and work with AI stakeholders to ensure that appropriately vetted datasets and models that are released for sharing are ready and fit for use and that they are maintained as standards and norms evolve. Ultimately, development and adoption of best practices and standards in documenting dataset and model provenance will enhance trustworthiness and responsible use of AI technologies.

Since 2016, there have also been increased concerns about data content, such as potential bias (see Strategy 3)[86,87] or private information leakage. The 2016 *National AI R&D Strategic Plan* noted that "dataset development and sharing must … follow applicable laws and regulations and be carried out in an ethical manner." The DOT-supported InSight project provides such carefully structured access to data collected during the Naturalistic Driving Study (see sidebar). The 2016 *National AI R&D Strategic Plan* also noted that new "technologies are needed to ensure safe sharing of data, since data owners take on risk when sharing their data with the research community." For example, CryptoNets[88] allows neural networks to operate over encrypted data, ensuring that data remain confidential, because decryption keys are not needed in neural networks. Researchers have also begun developing new ML techniques that use a differential privacy framework to provide quantifiable privacy guarantees over the used data.[89] At the same time, privacy methods must remain sufficiently explainable and transparent to help researchers correct them and make them safe, efficient, and accurate. Furthermore, AI could reveal discoveries beyond the original or intended scope; therefore, researchers must remain cognizant of the potential dangers in access to data or discoveries by adversaries.

Data alone are of little use without the ability to bring computational resources to bear on large-scale public datasets. The importance of computational resources to AI R&D is called out in the 2019 *Executive Order on Maintaining American Leadership in Artificial Intelligence*:[1]

> The Secretaries of Defense, Commerce, Health and Human Services, and Energy, the Administrator of the National Aeronautics and Space Administration, and the Director of the National Science Foundation shall, to the extent appropriate and consistent with applicable law, prioritize the allocation of high-performance computing resources for AI-related applications through: (i) increased assignment of discretionary allocation of resources and resource reserves; or (ii) any other appropriate mechanisms.

---

[86] Emily M. Bender and Batya Friedman, "Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science," *Transactions of the Association for Computational Linguistics* 6 (2018):587-604.

[87] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science* 356(6334):183-186, 14 Apr 2017.

[88] Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, John Wernsing, "CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy," *2016 International Conference on Machine Learning* 48:201-210; http://proceedings.mlr.press/v48/.

[89] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, "Deep Learning with Differential Privacy," *23rd ACM Conference on Computer and Communications Security*," 2016: 308-318.

and:

> …the Select Committee, in coordination with the General Services Administration (GSA), shall submit a report to the President making recommendations on better enabling the use of cloud computing resources for federally funded AI R&D.

The need for computational capacity for many AI challenges has been increasing rapidly.[32] Federal funding may provide computational capabilities for Federally-funded research. Some companies and universities, however, may have additional computational demands. Overall, there is a national need to study and invest in shared computational resources to promote AI R&D.

The benefits of AI will continue to accrue, but only to the extent that training and testing resources for AI are developed and made available. The variety, depth, quality, and accuracy of training datasets and other resources significantly affects AI performance. Many different AI technologies require high-quality data for training and testing, as well as dynamic, interactive testbeds and simulation environments. More than just a technical question, this is a significant "public good" challenge, as progress would suffer if AI training and testing is limited to only a few entities that already hold valuable datasets and resources, yet we must simultaneously respect commercial and individual rights and interests in the data. Research is needed to develop high-quality datasets and environments for a wide variety of AI applications and to enable responsible access to good datasets and testing and training resources. Additional open-source software libraries and toolkits are also needed to accelerate the advancement of AI R&D. The following subsections outline these key areas of importance.

Developing and making accessible a wide variety of datasets to meet the needs of a diverse spectrum of AI interests and applications

The integrity and availability of AI training and testing datasets is crucial to ensuring scientifically reliable results. The technical as well as the socio-technical infrastructure necessary to support reproducible research in the digital area has been recognized as an important challenge—and is essential to AI technologies as well. The lack of vetted and openly available datasets with identified provenance to enable reproducibility is a critical factor to confident advancement in AI.[90] As in other data-intensive sciences, capturing data provenance is critical. Researchers must be able to reproduce results with the same as well as different datasets. Datasets must be representative of challenging real-world applications, and not just simplified versions. To make progress quickly, emphasis should be placed on making available already existing datasets held by government, those that can be developed with Federal funding, and, to the extent possible, those held by industry.

The machine learning aspect of the AI challenge is often linked with "big data" analysis. Considering the wide variety of relevant datasets, it remains a growing challenge to have appropriate representation, access, and analysis of unstructured or semi-structured data. How can the data be represented—in absolute as well as relative (context-dependent) terms? Current real-world databases can be highly susceptible to inconsistent, incomplete, and noisy data. Therefore, a number of data preprocessing techniques (e.g., data cleaning, integration, transformation, reduction, and representation) are important to establishing useful datasets for AI applications. How does the data preprocessing impact data quality, especially when additional analysis is performed?

[90] Toward this end, in 2016 the Intelligence Advanced Research Projects Activity issued a Request for Information on novel training datasets and environments to advance AI. See https://iarpa.gov/index.php/working-with-iarpa/requests-for-information/novel-training-datasets-and-environments-to-advance-artificial-intelligence.

Encouraging the sharing of AI datasets—especially for government-funded research—would likely stimulate innovative AI approaches and solutions. However, technologies are needed to ensure safe sharing of data, since data owners take on risk when sharing their data with the research community. Dataset development and sharing must also follow applicable laws and regulations and be carried out in an ethical manner. Risks can arise in various ways: inappropriate use of datasets, inaccurate or inappropriate disclosure, and limitations in data de-identification techniques to ensure privacy and confidentiality protections.

Making training and testing resources responsive to commercial and public interests

With the continuing explosion of data, data sources, and information technology worldwide, both the number and size of datasets are increasing. The techniques and technologies to analyze data are not keeping up with the high volume of raw information sources. Data capture, curation, analysis, and visualization are all key research challenges, and the science needed to extract valuable knowledge from enormous amounts of data is lagging behind. While data repositories exist, they are often unable to deal with the scaling up of datasets, have limited data provenance information, and do not support semantically rich data searches. Dynamic, agile repositories are needed.

One example of the kind of open/sharing infrastructure program that is needed to support the needs of AI research is the IMPACT program (Information Marketplace for Policy and Analysis of Cyber-risk & Trust) developed by the Department of Homeland Security (DHS).[91] This program supports the global cyber security risk research effort by coordinating and developing real-world data and information sharing capabilities, including tools, models, and methodologies. IMPACT also supports empirical data sharing between the international cybersecurity R&D community, critical infrastructure providers, and their government supporters. AI R&D would benefit from comparable programs across all AI applications.

Developing open-source software libraries and toolkits

The increased availability of open-source software libraries and toolkits provides access to cutting-edge AI technologies for any developer with an Internet connection. Resources such as the Weka toolkit,[92] MALLET,[93] and OpenNLP,[94] among many others, have accelerated the development and application of AI. Development tools, including free or low-cost code repository and version control systems, as well as free or low-cost development languages (e.g., R, Octave, and Python) provide low barriers to using and extending these libraries. In addition, for those who may not want to integrate these libraries directly, any number of cloud-based machine learning services exist that can perform tasks such as image classification on demand through low-latency web protocols that require little or no programming for use. Finally, many of these web services also offer the use of specialized hardware, including GPU-based systems. It is reasonable to assume that specialized hardware for AI algorithms, including neuromorphic processors, will also become widely available through these services.

Together, these resources provide an AI technology infrastructure that encourages marketplace innovation by allowing entrepreneurs to develop solutions that solve narrow domain problems without requiring expensive hardware or software, without requiring a high level of AI expertise, and permitting rapid scaling-up of systems on demand. For narrow AI domains, barriers to marketplace innovation are extremely low relative to many other technology areas.

[91] https://www.dhs.gov/csd-impact

[92] https://sourceforge.net/projects/weka/

[93] http://mallet.cs.umass.edu

[94] https://opennlp.apache.org

To help support a continued high level of innovation in this area, the U.S. Government can boost efforts in the development, support, and use of open AI technologies. Particularly beneficial would be open resources that use standardized or open formats and open standards for representing semantic information, including domain ontologies when available.

Government may also encourage greater adoption of open AI resources by accelerating the use of open AI technologies within the government itself, and thus help to maintain a low barrier to entry for innovators. Whenever possible, government should contribute algorithms and software to open source projects. Because government has specific concerns, such as a greater emphasis on data privacy and security, it may be necessary for the government to develop mechanisms to ease government adoption of AI systems. For example, it may be useful to create a task force that can perform a "horizon scan" across government agencies to find particular AI application areas within departments, and then determine specific concerns that would need to be addressed to permit adoption of such techniques by these agencies.

Strategy 6: Measure and Evaluate AI Technologies through Standards and Benchmarks

2019 Update	**Supporting development of AI technical standards and related tools**

The 2016 *National AI R&D Strategic Plan* states that "Standards, benchmarks, testbeds, and their adoption by the AI community are essential for guiding and promoting R&D of AI technologies." In the intervening three years, emphasis on standards and benchmarks has continued to rise in the U.S. and globally. The 2019 *Executive Order on Maintaining American Leadership in Artificial Intelligence* explicitly calls out the importance of such standards:[1]

> …[T]he Secretary of Commerce, through the Director of [NIST], shall issue a plan for Federal engagement in the development of technical standards and related tools in support of reliable, robust, and trustworthy systems that use AI technologies.

With AI innovation potentially impacting all sectors and domains of society, many standards development organizations have new AI-related considerations and work items underway, including activities related to AI ethics and trustworthy AI systems (see Strategy 3). The International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) have convened a joint technical subcommittee on AI (ISO/IEC Joint Technical Committee 1, Subcommittee 42 on Artificial Intelligence[95]) to develop standards for AI systems and associated considerations. It is critical that Federal, industry, and academic researchers continue to inform these activities, particularly as AI advances and systems reach into areas such as transportation, health care, and food that align with the missions of government agencies.

Standards, benchmarks, and related tools: Recent agency R&D programs

Since the release of the 2016 *National AI R&D Strategic Plan*, NIST in particular has initiated efforts for Strategy 6:

- NIST is engaged in the standardization program of ISO/IEC JTC 1 SC 42 on Artificial Intelligence.[95] A NIST expert is the convener for the Big Data work effort in SC 42. The U.S. delegation to SC 42 includes NIST and other Federal agency experts, as well as representatives from industry and academia. U.S. input to SC 42 is facilitated by the International Committee for Information Technology Standards (INCITS).
- NIST staff participate in additional AI standards activities through standards organizations, such as the American Society of Mechanical Engineers, IEEE, and ISO/IEC. Their activities cover such topics as computational modeling for advanced manufacturing, ontologies for robotics and automation, personal data privacy, and algorithmic bias.
- NIST experts are raising awareness about the importance of consensus standards for AI in multilateral fora, including bodies such as G20 and G7.[96] NIST brings unique Federal Government expertise that grounds policy discussions in practice, in particular, through close collaboration with the private sector. Similarly, NIST lends its standards and related experience to intergovernmental bilateral discussions.

Since 2016, the surge in AI-related standards activities has outpaced the launch of new AI-focused benchmarks and evaluations, particularly as related to trustworthiness of AI systems. In the

[95] https://www.iso.org/committee/6794475.html

[96] https://home.treasury.gov/policy-issues/international/g-7-and-g-20

intervening time, however, considerations of fairness and bias in benchmark datasets have become increasingly important, with researchers pursuing new facial recognition datasets that seek to minimize bias. Much more plentiful are benchmarks that test the application-level performance of AI algorithms (e.g., false-positive or false-negative rates for classification algorithms) and benchmarks that quantify the compute-level performance of AI software and hardware systems. Two such recent activities are MLPerf[97] and DAWNbench.[98]

Assessing, promoting, and assuring all aspects of AI trustworthiness requires measuring and evaluating AI technology performance through benchmarks and standards. Beyond being safe, secure, reliable, resilient, explainable, and transparent, trustworthy AI must preserve privacy while detecting and avoiding inappropriate bias. As AI technologies evolve, so will the need to develop new metrics and testing requirements for validation of these essential characteristics.

Standards, benchmarks, testbeds, and their adoption by the AI community are essential for guiding and promoting R&D of AI technologies. The following subsections outline areas where additional progress must be made.

Developing a broad spectrum of AI standards

The development of standards must be hastened to keep pace with the rapidly evolving capabilities and expanding domains of AI applications. Standards provide requirements, specifications, guidelines, or characteristics that can be used consistently to ensure that AI technologies meet critical objectives for functionality and interoperability, and that they perform reliably and safely. Adoption of standards brings credibility to technology advancements and facilitates an expanded interoperable marketplace. One example of an AI-relevant standard that has been developed is P1872-2015 (Standard Ontologies for Robotics and Automation), developed by the Institute of Electrical and Electronics Engineers. This standard provides a systematic way of representing knowledge and a common set of terms and definitions. These allow for unambiguous knowledge transfer among humans, robots, and other artificial systems, as well as provide a foundational basis for the application of AI technologies to robotics. Additional work in AI standards development is needed across all subdomains of AI.

Standards are needed to address:

- *Software engineering*: to manage system complexity, sustainment, security, and to monitor and control emergent behaviors;
- *Performance*: to ensure accuracy, reliability, robustness, accessibility, and scalability;
- *Metrics*: to quantify factors impacting performance and compliance to standards;
- *Safety*: to evaluate risk management and hazard analysis of systems, human computer interactions, control systems, and regulatory compliance;
- *Usability*: to ensure that interfaces and controls are effective, efficient, and intuitive;
- *Interoperability*: to define interchangeable components, data, and transaction models via standard and compatible interfaces;
- *Security*: to address the confidentiality, integrity, and availability of information, as well as cybersecurity;
- *Privacy*: to control for the protection of information while being processed, when in transit, or being stored;

[97] https://mlperf.org/

[98] https://dawn.cs.stanford.edu/benchmark/

- *Traceability*: to provide a record of events (their implementation, testing, and completion), and for the curation of data; and
- *Domains*: to define domain-specific standard lexicons and corresponding frameworks.

Establishing AI technology benchmarks

Benchmarks, made up of tests and evaluations, provide quantitative measures for developing standards and assessing compliance to standards. Benchmarks drive innovation by promoting advancements aimed at addressing strategically selected scenarios; they additionally provide objective data to track the evolution of AI science and technologies. To effectively evaluate AI technologies, relevant and effective testing methodologies and metrics must be developed and standardized. Standard testing methods will prescribe protocols and procedures for assessing, comparing, and managing the performance of AI technologies. Standard metrics are needed to define quantifiable measures in order to characterize AI technologies, including but not limited to: accuracy, complexity, trust and competency, risk and uncertainty, explainability, unintended bias, comparison to human performance, and economic impact. It is important to note that benchmarks are data driven. Strategy 5 discusses the importance of datasets for training and testing.

As a successful example of AI-relevant benchmarks, the National Institute of Standards and Technology has developed a comprehensive set of standard test methods and associated performance metrics to assess key capabilities of emergency response robots. The objective is to facilitate quantitative comparisons of different robot models by making use of statistically significant data on robot capabilities that was captured using the standard test methods. These comparisons can guide purchasing decisions and help developers to understand deployment capabilities. The resulting test methods are being standardized though the ASTM International Standards Committee on Homeland Security Applications for robotic operational equipment (referred to as standard E54.08.01).[99] Versions of the test methods are used to challenge the research community through the RoboCup Rescue Robot League competitions,[100] which emphasize autonomous capabilities. Another example is the IEEE Agile Robotics for Industrial Automation Competition (ARIAC),[101] a joint effort between IEEE and NIST,[102] which promotes robot agility by utilizing the latest advances in artificial intelligence and robot planning. A core focus of this competition is to test the agility of industrial robot systems, with the goal of enabling those on the shop floors to be more productive, more autonomous, and requiring less time from shop floor workers.

While these efforts provide a strong foundation for driving AI benchmarking forward, they are limited by being domain-specific. Additional standards, testbeds, and benchmarks are needed across a broader range of domains to ensure that AI solutions are broadly applicable and widely adopted.

Increasing the availability of AI testbeds

The importance of testbeds was stated in the *Cyber Experimentation of the Future* report: "Testbeds are essential so that researchers can use actual operational data to model and run experiments on real-world system[s] … and scenarios in good test environments."[103] Having adequate testbeds is a

[99] 2019 update: The resulting test methods are now standards issued by ASTM International Standards Committee on Homeland Secruity Applications for Response Robots (referred to as E54.09).

[100] http://www.robocup2016.org/en/

[101] http://robotagility.wixsite.com/competition

[102] 2019 update: IEEE is no longer a partner of ARIAC, which is now in its third year.

[103] SRI International and USC Information Sciences Institute, "Cybersecurity Experimentation of the Future (CEF): Catalyzing a New Generation of Experimental Cybersecurity Research," Final Report, July 31, 2015.

need across all areas of AI. The government has massive amounts of mission-sensitive data unique to government, but much of this data cannot be distributed to the outside research community. Appropriate programs could be established for academic and industrial researchers to conduct research within secured and curated testbed environments established by specific agencies. AI models and experimental methods could be shared and validated by the research community by having access to these test environments, affording AI scientists, engineers, and students unique research opportunities not otherwise available.

Engaging the AI community in standards and benchmarks

Government leadership and coordination is needed to drive standardization and encourage its widespread use in government, academia, and industry. The AI community—made up of users, industry, academia, and government—must be energized to participate in developing standards and benchmark programs. As each government agency engages the community in different ways based on its role and mission, community interactions can be leveraged through coordination in order to strengthen their impact. This coordination is needed to collectively gather user-driven requirements, anticipate developer-driven standards, and promote educational opportunities. User-driven requirements shape the objectives and design of challenge problems and enable technology evaluation. Having community benchmarks focuses R&D to define progress, close gaps, and drive innovative solutions for specific problems. These benchmarks must include methods for defining and assigning ground truth. The creation of benchmark simulation and analysis tools will also accelerate AI developments. The results of these benchmarks also help match the right technology to the user's need, forming objective criteria for standards compliance, qualified product lists, and potential source selection.

Industry and academia are the primary sources for emerging AI technologies. Promoting and coordinating their participation in standards and benchmarking activities are critical. As solutions emerge, opportunities abound for anticipating developer- and user-driven standards through sharing common visions for technical architectures, developing reference implementations of emerging standards to show feasibility, and conducting precompetitive testing to ensure high-quality and interoperable solutions, as well as to develop best practices for technology applications.

One successful example of a high-impact, community-based, AI-relevant benchmark program is the Text Retrieval Conference (TREC),[104] which was started by NIST in 1992 to provide the infrastructure necessary for large-scale evaluation of information retrieval methodologies. More than 250 groups have participated in TREC, including academic and commercial organizations both large and small. The standard, widely available, and carefully constructed set of data put forth by TREC has been credited with revitalizing research on information retrieval.[105,106] A second example is the NIST periodic benchmark program in the area of machine vision applied to biometrics,[107] particularly face recognition.[108] This began with the Face Recognition Technology (FERET) evaluation in 1993, which provided a standard dataset of face photos designed to support face recognition algorithm development as well as an evaluation protocol. This effort has evolved over the years into the Face

[104] http://trec.nist.gov

[105] E. M. Voorhees and D. K. Harman, *TREC Experiment and Evaluation in Information Retrieval* (Cambridge: MIT Press, 2005).

[106] http://googleblog.blogspot.com/2008/03/why-data-matters.html

[107] http://biometrics.nist.gov

[108] http://face.nist.gov

Recognition Vendor Test (FRVT),[109] involving the distribution of datasets, hosting of challenge problems, and conducting of sequestered technology evaluations. This benchmark program has contributed greatly to the improvement of facial recognition technology. Both TREC and FRVT can serve as examples of effective AI-relevant community benchmarking activities, but similar efforts are needed in other areas of AI.

It is important to note that developing and adopting standards, as well as participating in benchmark activities, comes with a cost. R&D organizations are incentivized when they see significant benefit. Updating acquisition processes across agencies to include specific requirements for AI standards in requests for proposals will encourage the community to further engage in standards development and adoption. Community-based benchmarks, such as TREC and FRVT, also lower barriers and strengthen incentives by providing types of training and testing data otherwise inaccessible, fostering healthy competition between technology developers to drive best-of-breed algorithms, and providing objective and comparative performance metrics for relevant source selections.

[109] P. J. Phillips, "Improving Face Recognition Technology," *Computer* 44(3)(2011): 84-96.

Strategy 7: Better Understand the National AI R&D Workforce Needs

2019 Update	**Advancing the AI R&D workforce, including those working on AI systems and those working alongside them, to sustain U.S. leadership**

Since the release of the 2016 *National AI R&D Strategic Plan*, the demand for AI researchers and practitioners has grown rapidly. Studies have shown that the number of hiring opportunities is expected to rise into the millions over the next decade. As one data point, the U.S. Bureau of Labor Statistics projects that the number of positions for computer and information scientists and engineers will grow by 19% from 2016 to 2026, almost three times faster than the average for all occupations.[111] Moreover, through 2028, AI researchers are expected contribute to as much as $11.5 trillion of cumulative growth promised by intelligent technologies in the G20 countries alone.[112]

U.S. academic institutions are struggling to keep pace with the explosive growth in student interest and enrollment in AI.[113,114,115] At the same time, industry, with its sustained financial support and access to advanced computing facilities and datasets, exerts a strong pull on academic research and teaching talent.[116]

It is critical to maintain a robust academic research ecosystem in AI that, in collaboration with industry R&D, can continue to deliver tremendous dividends[117] by advancing national health, prosperity, and welfare, and securing the national defense.

National AI R&D workforce: Recent agency activities

Since the release of the 2016 *National AI R&D Strategic Plan*, a number of agencies have initiated efforts supporting Strategy 7:

- Apart from supporting undergraduate and graduate students through standard AI research grants, agencies are prioritizing computational- and data-enabled science and engineering in their graduate fellowship programs. For example, in 2018, DOE added a new track to its Computational Science Graduate Fellowship program. This track supports students pursuing advanced degrees in applied mathematics, statistics, or computer science, and promotes more effective use of high-performance systems, including in the areas of AI, ML, and deep learning.[44,110] Also in 2018, NSF began prioritizing computational and data-enabled science and engineering in a subset of awardees of its Graduate Research Fellowships Program.
- The Census Bureau has created the Statistical Data Modernization (SDM) project to bring its workforce, operations, and technologies up to the current state of the art and set the standard for statistical agencies in today's data-driven society. SDM's workforce transformation component will enable the hiring of new data scientists with expertise in new methods and analytics, including the use of AI methods and tools to process and analyze big data. The workforce transformation will also address the upskilling of the current data science workforce.

[110] https://www.krellinst.org/csgf/math-cs

[111] https://www.bls.gov/ooh/computer-and-information-technology/computer-and-information-research-scientists.htm

[112] https://www.accenture.com/t20180920T094705Z__w__/us-en/_acnmedia/Thought-Leadership-Assets/PDF/Accenture-Education-and-Technology-Skills-Research.pdf

[113] https://cra.org/data/generation-cs/

[114] https://cra.org/wp-content/uploads/2018/05/2017-Taulbee-Survey-Report.pdf

[115] http://web.cs.wpi.edu/~cew/papers/CSareas19.pdf

[116] https://www.nitrd.gov/rfi/ai/2018/AI-RFI-Response-2018-Yolanda-Gil-AAAI.pdf

[117] https://www.nap.edu/catalog/13427/continuing-innovation-in-information-technology

In the three years since the release of the 2016 *National AI R&D Strategic Plan*, various reports have called for continued support for the development of instructional materials and teacher professional development in computer science at all levels. Emphasis is needed at the K–12 levels to feed the Nation's pipeline of AI researchers over many decades.[118] At the undergraduate level, there is a need to focus on integrating advanced computational skills and methods with domain-specific knowledge from other disciplines, given the growing role of computing across disciplines.[119] Sustained support is also needed at the graduate level, where students are conducting fundamental research in ML and AI. Indeed, the 2019 *Executive Order on Maintaining American Leadership in Artificial Intelligence* requires that:[1]

> Heads of implementing agencies that also provide educational grants shall, to the extent consistent with applicable law, consider AI as a priority area within existing Federal fellowship and service programs … [including] … (A) high school, undergraduate, and graduate fellowship; alternative education; and training programs; (B) programs to recognize and fund early-career university faculty who conduct AI R&D, including through Presidential awards and recognitions; (C) scholarship for service programs; (D) direct commissioning programs of the United States Armed Forces; and (E) programs that support the development of instructional programs and curricula that encourage the integration of AI technologies into courses in order to facilitate personalized and adaptive learning experiences for formal and informal education and training.

More broadly, the need for a firm grounding in computational thinking, including through computer science education, is also noted prominently in the Federal Government's December 2018 five-year strategic plan for science, technology, engineering, and mathematics (STEM) education.[120]

In addition, it is imperative to broaden the participation among groups traditionally underrepresented in computing and related fields.

Finally, the AI R&D workforce will consist of multidisciplinary teams comprising not just computer and information scientists and engineers, but also experts from other fields key to AI and ML innovation and its application, including cognitive science and psychology, economics and game theory, engineering and control theory, ethics, linguistics, mathematics, philosophy, and the many domains in which AI may be applied.

Federal agencies are giving priority to training and fellowship programs at all levels to prepare the workforce with requisite AI R&D skills through apprenticeships, skills programs, fellowships, and course work in relevant disciplines (see sidebar). Such training opportunities target both scientists and engineers who contribute to AI R&D innovations and users of AI R&D who may possess relevant domain knowledge. In the case of the former, long-term Federal investment in AI R&D, as described in Strategy 1, further supports the growth of this workforce, both through training the next generation of researchers and by making faculty positions more attractive to current graduate and postdoctoral students. In the case of the latter, new programs are bringing AI-relevant skills to current and future users of AI systems (see sidebar). Federal agencies must therefore continue to strategically foster expertise in the AI R&D workforce that spans multiple disciplines and skill categories to ensure sustained national leadership.

[118] https://github.com/touretzkyds/ai4k12/wiki

[119] https://www.nap.edu/catalog/24926/assessing-and-responding-to-the-growth-of-computer-science-undergraduate-enrollments

[120] https://www.whitehouse.gov/wp-content/uploads/2018/12/STEM-Education-Strategic-Plan-2018.pdf

Attaining the needed AI R&D advances outlined in this strategy will require a sufficient AI R&D workforce. Nations with the strongest presence in AI R&D will establish leading positions in the automation of the future. They will become the frontrunners in competencies like algorithm creation and development; capability demonstration; and commercialization. Developing technical expertise will provide the basis for these advancements.

While no official AI workforce data currently exist, numerous recent reports from the commercial and academic sectors are indicating an increased shortage of available experts in AI. AI experts are reportedly in short supply,[121] with demand expected to continue to escalate.[122] High-tech companies are reportedly investing significant resources into recruiting faculty members and students with AI expertise.[123] Universities and industries are reportedly in a battle to recruit and retain AI talent.[124]

Additional studies are needed to better understand the current and future national workforce needs for AI R&D. Data is needed to characterize the current state of the AI R&D workforce, including the needs of academia, government, and industry. Studies should explore the supply and demand forces in the AI workplace, to help predict future workforce needs. An understanding is needed of the projected AI R&D workforce pipeline. Considerations of educational pathways and potential retraining opportunities should be included. Diversity issues should also be explored, since studies have shown that a diverse information technology workforce can lead to improved outcomes.[125] Once the current and future AI R&D workforce needs are better understood, then appropriate plans and actions can be considered to address any existing or anticipated workforce challenges.

[121] "Startups Aim to Exploit a Deep-Learning Skills Gap," *MIT Technology Review*, January 6, 2016.

[122] "AI talent grab sparks excitement and concern," *Nature*, April 26, 2016.

[123] "Artificial Intelligence Experts are in High Demand," *The Wall Street Journal*, May 1, 2015.

[124] "Million dollar babies: As Silicon Valley fights for talent, universities struggle to hold on to their stars," *The Economist*, April 2, 2016.

[125] J. W. Moody, C. M. Beise, A. B. Woszczynski, and M. E. Myers, "Diversity and the information technology workforce: Barriers and opportunities," *Journal of Computer Information Systems* 43 (2003): 63-71.

Strategy 8: Expand Public–Private Partnerships to Accelerate Advances in AI

Strategy 8 is new in 2019 and reflects the growing importance of public-private partnerships enabling AI R&D.

American leadership in science and engineering research and innovation is rooted in the Nation's unique government-university-industry R&D ecosystem. As the American Association of Arts and Sciences has written, "America's standing as an innovation leader" relies upon "establishing a more robust national Government-University-Industry research partnership."[126] Since the release of the 2016 *National AI R&D Strategic Plan*, the Administration has amplified this vision of promoting "sustained investment in AI R&D in collaboration with academia, industry, international partners and allies, and other non-Federal entities to generate technological breakthroughs in AI and related technologies and to rapidly transition those breakthroughs into capabilities that contribute to U.S. economic and national security."[1]

Over the last several decades, fundamental research in information technology conducted at universities with Federal funding, as well as in industry, has led to new, multi-billion-dollar sectors of the Nation's economy.[127] Concurrent advances across government, universities, and industry have been mutually reinforcing and have led to an innovative, vibrant AI sector. Many of today's AI systems have been enabled by the American government-university-industry R&D ecosystem (see sidebar).

Advancing the Nation's AI innovation ecosystem, spanning government, universities, and industry

- Deep convolutional neural networks have proven to be a key innovation rooted in AI research. Although this modeling approach emerged from early Federal investments in the late 1980s, there were not enough data nor enough computational capabilities available at the time for neural networks to make accurate predictions. Through the combination of a rise in big data, today's data-intensive scientific methods, and conceptual advances in how to structure and optimize the networks, neural networks have re-emerged as a useful way to improve accuracy in AI models. Interactions between academia and the private sector, including government funding, in recent years have helped reduce the error rate in speech recognition systems, enabling innovations such as real-time translation.[126]
- Similarly, Federal investments in reinforcement learning in the 1980s and 1990s—an approach rooted in behavioral psychology that involves learning to associate behaviors with desired outcomes—have led to today's deep learning systems. Through interactions across sectors, computers are increasingly learning like humans, without explicit instruction, and reinforcement learning is driving progress in self-driving cars and other forms of automation where machines can hone skills through experience. Reinforcement learning was the key technology underlying AlphaGo, the program that defeated the world's best Go players, which has seen a growing number of victories over professional players since 2016.[126]

Since the release of the 2016 *National AI R&D Strategic Plan*, additional emphasis has been placed on the benefits of public-private partnerships. These benefits include strategically leveraging resources, including facilities, datasets, and expertise, to advance science and engineering innovations;

[126] *Restoring the Foundation: The Vital Role of Research in Preserving the American Dream* (American Academy of Arts and Sciences, Cambridge, MA, 2014); https://www.amacad.org/multimedia/pdfs/publications/researchpapersmonographs/AmericanAcad_RestoringtheFoundation_Brief.pdf .

[127] National Research Council Computer Science Telecommunications Board, *Continuing Innovation in Information Technology* (The National Academies Press, Washington, D.C., 2012); https://doi.org/10.17226/13427.

accelerating the transition of these innovations to practice; and enhancing education and training for next-generation researchers, technicians, and leaders. Government-university-industry R&D partnerships bring pressing, real-world challenges faced by industry to university researchers, enabling "use-inspired research"; leverage industry expertise to accelerate the transition of open and published research results into viable products and services in the marketplace for economic growth; and grow research and workforce capacity by linking university faculty and students with industry representatives, industry settings, and industry jobs (see sidebar).[126,128,129,130] These partnerships build upon joint engagements among Federal agencies that enable synergies in areas where agencies' missions intersect. The Nation also benefits from relationships between Federal agencies and international funders who can work together to address key challenges of mutual interest across a range of disciplines.

While there are many structures and mechanisms for public-private partnerships, some common categories for engagement include:

1. *Individual project-based collaborations.* These efforts enable engagement by university researchers with those in other sectors, including Federal agencies, industry, and international organizations, to identify and leverage synergies in areas of mutual interest.
2. *Joint programs to advance open, precompetitive, fundamental research*. Direct partnerships among organizations across sectors enable funding and support for open, precompetitive, fundamental research in areas of mutual interest to the partners. In general, non-Federal partners contributing research resources receive the same intellectual property rights afforded to the U.S. Government by the Bayh-Dole Act.[131]
3. *Collaborations to deploy and enhance research infrastructure*. Collaborations among Federal agencies, industry, and international organizations significantly enhance the potential for developing new and enhancing existing research infrastructure that in turn enables groundbreaking experimentation by researchers. Partners may offer financial and/or in-kind support to develop and/or enhance research infrastructure.
4. *Collaborations to enhance workforce development including broadening participation*. Multisector partnerships set the foundation for rigorous, engaging, and inspiring instructional materials that enhance workforce development and diversity in STEM professions.

In each of these cases, leveraging each partner's strengths for the benefit of all is vitally important to achieving success.

[128] Mathematical Sciences Research Institute report, "*Partnerships: A Workshop on Collaborations between the NSF/MPS & Private Foundations,"* 2015; http://library.msri.org/msri/Partnerships.pdf.

[129] Computing Community Consortium, *"The Future of Computing Research: Industry-Academic Collaborations,"* 2016; http://cra.org/ccc/wp-content/uploads/sites/2/2016/06/15125-CCC-Industry-Whitepaper-v4-1.pdf.

[130] Computing Community Consortium, *"Evolving Academia/Industry Relations in Computing Research: Interim Report released by the CCC,"* 2019; https://www.cccblog.org/wp-content/uploads/2019/03/Industry-Interim-Report-w-footnotes.pdf.

[131] https://history.nih.gov/research/downloads/PL96-517.pdf

Advances in AI R&D stand to benefit from all of these types of public-private partnerships. Partnerships can promote open, precompetitive, fundamental AI R&D; enhance access to research resources such as datasets, models, and advanced computational capabilities; and foster researcher exchanges and/or joint appointments between government, universities, and industry to share AI R&D expertise. Partnerships can also promulgate the inherently interdisciplinary nature of AI R&D, which requires convergence between computer and information science, cognitive science and psychology, economics and game theory, engineering and control theory, ethics, linguistics, mathematics and statistics, and philosophy to drive the development and evaluation of future AI systems that are fair, transparent, and accountable, as well as safe and secure. Federal agencies are actively pursuing public-private partnerships to achieve these goals (*see sidebar*).

Federal agencies must therefore continue to pursue and strengthen public-private partnerships in AI R&D to drive technology development and economic growth by leveraging investments and expertise in areas of mutual interest to government, industry, and academia. In doing so, the U.S. Government will capitalize on a uniquely American innovation ecosystem that has transformed nearly every aspect of the Nation's economy and society over the last two decades through novel information technologies.[127]

Public-private partnerships: Recent agency R&D programs

A number of agencies have already initiated public-private partnerships in support of AI R&D:

- The Defense Innovation Unit (DIU)[132] is a DoD organization that solicits commercial solutions capable of addressing DoD needs. The DIU in turn provides pilot contracts, which can include hardware, software, or other unique services. If successful, pilot contracts lead to follow-on contracts between companies and any DoD entity. A key DIU feature is the rapid pace of the pilot and subsequent contracts.
- NSF and the Partnership on AI, a diverse, multistakeholder organization working to better understand AI's impacts, are partnering to jointly support high-risk, high-reward research at the intersection of the social and technical dimensions of AI.[15]
- The DHS Science and Technology Directorate's Silicon Valley Innovation Program (SVIP)[133] looks to harness commercial R&D innovation ecosystems across the Nation and around the world for technologies with government applications. SVIP employs a streamlined application and pitch process; brings government, entrepreneurs, and industry together to find cutting-edge solutions; and co-invests in and accelerates transition to market.
- The Department of Health and Human Services (HHS) piloted the Health Tech Sprint initiative, also known in its first iteration as "Top Health," modeled in part after the Census Bureau's Opportunity Project. This effort created a nimble framework to public-private collaborations around bidirectional data links. It piloted new models for iterating on data release for AI training and testing, and it developed a voluntary incentivization framework for a public–private AI ecosystem.
- The HHS Division of Research, Innovation, and Ventures is part of the Biomedical Advanced Research and Development Authority at the Office of the Assistant Secretary for Preparedness and Response. It oversees an accelerator network and is recruiting a nonprofit partner that can work with private investors to fund innovative technologies and products to solve systemic health security challenges, with AI applications being one area of interest. Accelerators will connect startups and other businesses with product development and business support services.

[132] https://www.diu.mil/

[133] https://www.dhs.gov/science-and-technology/svip

Abbreviations

AFOSR	Air Force Office of Scientific Research
AI	artificial intelligence
DARPA	Defense Advanced Research Projects Agency
DHS	Department of Homeland Security
DoD	Department of Defense
DOE	Department of Energy
DOT	Department of Transportation
FDA	Food and Drug Administration
FRVT	Face Recognition Vendor Test
GPS	Global Positioning System
GPU	graphics processing unit
GSA	General Services Administration
HHS	Department of Health and Human Services
HPC	high-performance computing
IARPA	Intelligence Advanced Research Projects Activity
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IMPACT	Information Marketplace for Policy and Analysis of Cyber-risk & Trust (DHS)
ISO	International Organization for Standardization
IT	information technology
IWG	interagency working group
ML	machine learning
MLAI	Machine Learning and Artificial Intelligence (Subcommittee of the NSTC)
NASA	National Aeronautics and Space Administration
NCO	National Coordination Office for NITRD
NDS	Naturalistic Driving Study (DOT)
NIFA	National Institute of Food and Agriculture (USDA)
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NITRD	Networking and Information Technology Research and Development program
NLM	National Library of Medicine (NIH)
NSF	National Science Foundation
NSTC	National Science and Technology Council
NTIA	National Telecommunications and Information Administration
ODNI	Office of the Director of National Intelligence
OSTP	Office of Science and Technology Policy
R&D	research and development
RFI	Request for Information
STEM	science, technology, engineering, and mathematics
SVIP	Silicon Valley Innovation Program (DHS)
TREC	Text Retrieval Conference
USDA	U.S. Department of Agriculture
VA	U.S. Department of Veterans Affairs
XAI	explainable AI

National Science & Technology Council

Chair

Kelvin Droegemeier, Director, OSTP

Staff

Chloé Kontos, Executive Director, NSTC

Select Committee on Artificial Intelligence

Co-Chairs

Michael Kratsios, Deputy Assistant to the President for Technology Policy (The White House)

France A. Córdova, Director, NSF

Steven Walker, Director, DARPA

Subcommittee on Machine Learning and Artificial Intelligence

Co-Chairs

Lynne Parker, Assistant Director for Artificial Intelligence, OSTP

Charles Romine, Director, Information Technology Laboratory, NIST

James Kurose, Assistant Director, Directorate for Computer Information Science and Engineering (CISE), NSF

Stephen Binkley, Deputy Director, Science Programs, Office of Science, DOE

Executive Secretary

Faisal D'Souza, NITRD NCO

Subcommittee on Networking & Information Technology Research & Development

Co-Chairs

Kamie Roberts, Director, NITRD NCO

James Kurose, Assistant Director, CISE, NSF

Executive Secretary

Nekeia Butler, NITRD NCO

Artificial Intelligence Research & Development Interagency Working Group

Co-Chairs

Jeff Alstott, Program Manager, IARPA Office of the Director of National Intelligence (ODNI)

Henry Kautz, Division Director, CISE Division of Information and Intelligent Systems, NSF

Staff

Faisal D'Souza, Technical Coordinator, NITRD NCO

Strategic Plan Writing Team

Jeff Alstott, IARPA
Gil Alterovitz, VA
Sameer Antani, NIH
Charlotte Baer, NIFA, USDA
Daniel Clouse, ODNI
Faisal D'Souza, NITRD NCO
Kimberly Ferguson-Walter, U.S. Navy
Michael Garris, NIST
Erwin Gianchandani, NSF
Ross Gillfillan, OSTP
Travis Hall, NTIA
Meghan Houghton, NSF
Henry Kautz, NSF
Erin Kenneally, DHS
David Kuehn, DOT
James Kurose, NSF
James Lawton, AFOSR
Steven Lee, DOE
Aaron Mannes, DHS
Lynne Parker, OSTP
Dinesh Patwardhan, FDA
Elham Tabassi, NIST

About the National Science and Technology Council

The National Science and Technology Council (NSTC) is the principal means by which the Executive Branch coordinates science and technology policy across the diverse entities that make up the Federal research and development enterprise. A primary objective of the NSTC is to ensure that science and technology policy decisions and programs are consistent with the President's stated goals. The NSTC prepares research and development strategies that are coordinated across Federal agencies aimed at accomplishing multiple national goals. The work of the NSTC is organized under committees that oversee subcommittees and working groups focused on different aspects of science and technology. More information is available at https://www.whitehouse.gov/ostp/nstc.

About the Office of Science and Technology Policy

The Office of Science and Technology Policy (OSTP) was established by the National Science and Technology Policy, Organization, and Priorities Act of 1976 to provide the President and others within the Executive Office of the President with advice on the scientific, engineering, and technological aspects of the economy, national security, homeland security, health, foreign relations, the environment, and the technological recovery and use of resources, among other topics. OSTP leads interagency science and technology policy coordination efforts, assists the Office of Management and Budget with an annual review and analysis of Federal research and development (R&D) in budgets, and serves as a source of scientific and technological analysis and judgment for the President with respect to major policies, plans, and programs of the Federal Government. More information is available at https://www.whitehouse.gov/ostp.

About the Select Committee on Artificial Intelligence

The Select Committee on Artificial Intelligence (AI) advises and assists the NSTC to improve the overall effectiveness and productivity of Federal R&D efforts related to AI to ensure continued U.S. leadership in this field. It addresses national and international policy matters that cut across agency boundaries, and it provides formal mechanisms for interagency policy coordination and development for Federal AI R&D activities, including those related to autonomous systems, biometric identification, computer vision, human-computer interactions, machine learning, natural language processing, and robotics. It also advises the Executive Office of the President on interagency AI R&D priorities; works to create balanced and comprehensive AI R&D programs and partnerships; leverages Federal data and computational resources across department and agency missions; and supports a technical, national AI workforce.

About the Subcommittee on Machine Learning and Artificial Intelligence

The Machine Learning and Artificial Intelligence (MLAI) Subcommittee monitors the state of the art in machine learning (ML) and artificial intelligence within the Federal Government, in the private sector, and internationally to watch for the arrival of important technology milestones in the development of AI, to coordinate the use of and foster the sharing of knowledge and best practices about ML and AI by the Federal Government, and to consult in the development of Federal MLAI R&D priorities. The MLAI Subcommittee reports to the Committee on Technology and the Select Committee on AI. The MLAI Subcommittee also coordinates AI taskings with the Artificial Intelligence Research & Development Interagency Working Group (see below).

About the Subcommittee on Networking & Information Technology Research & Development

The Networking and Information Technology Research and Development (NITRD) Program is the Nation's primary source of Federally funded work on pioneering information technologies (IT) in computing, networking, and software. The NITRD Subcommittee guides the multiagency NITRD Program in its work to provide the R&D foundations for assuring continued U.S. technological leadership and meeting the needs of the Nation for advanced IT. It reports to the NSTC Committee on Science and Technology Enterprise. The Subcommittee is supported by the interagency working groups that report to it and by its National Coordination Office. More information is available at https://www.nitrd.gov/about/.

About the Artificial Intelligence Research & Development Interagency Working Group

The NITRD AI R&D Interagency Working Group (IWG) coordinates Federal R&D in AI; it also supports and coordinates activities tasked by the Select Committee on AI and the NSTC Subcommittee on Machine Learning and Artificial Intelligence. This vital work promotes U.S. leadership and global competitiveness in AI R&D. The NITRD AI R&D IWG spearheaded the update of this National Artificial Intelligence Research and Development Strategic Plan. More information is available at https://www.nitrd.gov/groups/AI.

About this Document

This document includes the original text from the 2016 *National AI R&D Strategic Plan* with updates prepared in 2019 following Administration and interagency evaluation of the 2016 Plan and of community responses to a Request for Information on updating the Plan. The 2016 strategies were broadly determined to be valid going forward with some reemphases and with a call for a new strategy on Private-Public Partnerships in AI. A shaded call-out box has been inserted at the top of each strategy to highlight updated imperatives and/or new focus areas. The 2019 update adds an entirely new Strategy 8 on Private-Public Partnerships in AI.

Copyright Information

Published in the United States of America, 2019.

THE WHITE HOUSE