CCID 赛迪智库 译丛

2017 年 1 月 16 日　　第 3 期　　总 288 第期

美国人工智能研究与发展战略计划

【译者按】为应对人工智能的发展，2016 年 10 月，奥巴马政府在白宫前沿峰会发布报告《美国人工智能研究与发展战略计划》，此计划明确了人工智能发展的七大战略，即对人工智能研发进行长期投资、开发人机协作的有效方法、积极应对人工智能带来的伦理、法律和社会影响、确保系统的安全性、开发人工智能训练和测试数据共享平台、建立人工智能标准评估体系和把握研发人才需求情况。此外还提出两条保障措施，一是制定人工智能研发实施框架，二是建立人工智能人才保障机制。赛迪智库规划研究所对该报告进行了编译，希望能为我国相关决策部门提供参考。

【关键词】美国 人工智能 发展战略 发展建议 研发框架

《美国人工智能研究与发展战略计划》（简称《战略计划》）的愿景是推动人工智能的快速发展，促进经济繁荣，增加教育机会，增强国家与国土安全。确定了美国联邦政府应对人工智能关键技术与社会挑战的短期与长期任务，以指导联邦政府顺利开展与人工智能有关的任务驱动型项目和投资计划。

战略一：对人工智能研发进行长期投资

需要对有潜在长期回报的人工智能研究领域进行投资，对高风险研究的持续投资可带来高回报。

（一）推动发展数据型知识挖掘方法

要做到理解数据和知识挖掘智能化，就需要使用许多基础的新工具与技术。一是开发更先进的、可以识别隐藏在大数据中的实用信息的机器学习算法。在处理大量数据时，数据的真实性是一大难题，这让人类难以从中评估和获取知识。要提高数据清理技术的效率、创建发现数据不一致与异常的方法以及设计处理人类反馈的方法，那么仍需要研究人员探索新的方法，从而实现数据与相关元数据的同步挖掘。二是许多人工智能应用都具有跨学科性并且用到了异构数据。为了实现对各种不同类型数据（比如离散、连续、文本、空间、临时、时空、图形）的知识挖掘，仍需要对多模态机器学习进行进一步的研究。

（二）增强广义人工智能系统的感知能力

感知能力是智能系统了解这个世界的窗口，感知来源于（可能是分布式）传感器数据，传感器数据常常连同先验知识和模型被一同处理并融合，从而获取与人工智能系统任务相关的信息。整合后的感知数据形成了态势感知，提供了周围世界综合知识和状态模型，是人工智能系统安全有效地规划和执行各项工作所必需的。传感器必须能够在长距离以更高的分辨率实时捕捉数据。感知系统需要能够整合来自各种传感器和计算云等其他来源的数据，确定人工智能系统当前的感知内容并且对未来状态进行预测。此外，为了量化人工智能系统态势感知的可信度以及提高准确性，需要设计一个在整个感知过程中计算并降低不确定性的架构。

（三）了解人工智能的理论能力与局限性

目前，我们还没有很好地了解人工智能的理论能力与局限以及人工智能算法实现类人解决方案的可能性。为了更好地了解人工智能技术，尤其是机器学习为何能够在实践中具有出色表现，我们需要对理论展开研究。虽然各学科（包括数学、控制科学以及计算机科学）都在研究这一问题，但该领域目前缺乏统一的理论模型或框架来了解人工智能系统的性能。同时，也需要对计算可解性开展研究，从而了解人工智能算法理论上能够解决和无法

解决的问题。同时，必须在现有硬件的背景下挖掘这些知识，从而了解硬件如何影响这些算法的性能。

（四）对通用人工智能展开研究

人工智能可以分为狭义人工智能和广义人工智能。狭义人工智能系统执行特定细分领域的任务，比如语音识别、图像识别以及翻译，只能应用于所设计的任务。广义人工智能的长期目标是建立能够在广泛认知领域展现人类智力灵活性与多样性的系统，包括学习、语言、感知、推理、创造与规划等。广泛的学习能力将使广义人工智能系统能够将一个领域的知识运用于另一个领域，以及通过交互方式从经验中学习和向人类学习。广义人工智能一直是研究人员的一个远大目标，但当前的系统离这个目标还很远。

（五）开发可扩展的人工智能系统

人工智能系统组和网络可以通过协调或自动调节来执行单一人工智能系统无法完成的任务，并且人类也可参与任务或者领导团队。未来的研究必须探索更高效、稳健与可扩展的人工智能系统群组与人类构成团队的规划、控制与协作技术。

（六）推动类人人工智能领域的研究

如要实现类人人工智能，就需要用人类可以理解的方式解释自身的系统。这会产生新一代智能系统，比如能够有效帮助人类

完成工作的智能辅导系统与智能辅助系统。但当前人工智能算法的工作方式与人类的学习与工作方式存在较大差异，人类可以从少数几个例子中学习，也可以通过获取完成工作的指示和/或“提示”或通过观察其他人如何完成工作来学习。通过在更基础的层面研究实现类人人工智能的新方法可能会让这些系统更接近这个目标。

（七）开发更强大更可靠的机器人

尽管机器人曾经被设计用于静态工业环境，但近期却朝着和人类密切合作的方向发展。如今，机器人技术正在展现其补充、放大、增强或模拟人类身体能力或人类智能的能力。为了让机器人更好地理解并与现实世界互动，需要在认知与推理方面取得进展。当机器人的适应与学习能力提高后，就能归纳自己的技能、对当前表现进行自我评估并且从人类老师那里学到各种身体动作。科学家需要让这些机器人系统变得更强大、可靠和便于使用。

（八）推动人工智能硬件发展

尽管人工智能研究大部分与软件的发展相关，但人工智能系统的性能十分依赖于它所依附的硬件。目前的深度机器学习直接与图形处理器（GPU）硬件技术的发展及其存储器、输入/输出、时钟速度、并行性和能效的改进息息相关。

（九）创造提高硬件性能的人工智能

硬件性能的提高可以增强人工智能系统的能力，而人工系统也可以提高硬件的性能。这一互惠性将进一步推动硬件性能的提高，因为突破计算的物理局限就需要新的硬件设计方案。人工智能目前正被用于预测高性能计算的性能和资源的使用，以及做出提高效率的网络优化决策；更先进的人工智能技术可以进一步增强系统性能。人工智能还能被用于创建能够在系统发生故障时，无需人类干预而处理系统故障的自动重新配置型高性能计算机系统。现实中，高性能计算机系统中的执行配置始终是不同的，并且会同步执行不同的应用，各软件代码的状态也会适时发生独立的变化。人工智能算法需要能够在线运行并且符合高性能计算机系统的规模。

战略二：开发人机协作的有效方法

人类与人工智能系统的互补组合会比自主人工智能系统更有效地实现应用目标。目前的人机合作方式大多是适用特定目标、特定平台、特定环境并无法扩展的“点解决方案”。因此，未来应该设计使用面广、建立简便、切换快捷的通用系统，还是建立大量解决特定问题、针对性强的系统，是一个需要权衡的问题。在未来应用中，人机之间的职能分工、互动性质、参与数量，

以及如何沟通和分享态势感知将发生显著的变化。

人工智能系统在与人类的分工中分几种角色，一是执行人类决策者的外围支持任务；二是在人类需要协助时执行复杂的监控功能、决策制定以及自动医疗诊断；三是代替人类执行超出人类能力范围的工作，以及需要系统做出迅速反应的场景。要实现人类与人工智能系统之间的有效互动，就需要开展更多的研发工作来保证该系统设计不会导致过高的复杂性、不信任或过度信任。为了保证人类有效地理解人工智能系统，并明确其能力范围，人工智能系统的设计与开发采用以人类为中心的自动化原理：采用直观、用户友好的人机系统接口、控制与显示设计；为操作人员提供系统状态及变更信息；为操作人员提供通用知识、技术和能力（KSA）的反复训练以及人工系统算法和逻辑与系统预期故障模式培训；将人工智能系统的部署作为操作人员的设计选项，实现灵活的自动化。创建有效人机合作的人工智能系统面临许多基本的挑战。

（一）寻找人类感知的人工智能的新算法

人工智能算法的能力与人类可应用能力存在差距，我们需要能感知人类的智能系统与用户进行直观的浅层次、甚至意图深入模型的互动，并且实现人机默契合作。如果智能系统可以具备一定程度的情感智能，从而识别用户的情感并且做出恰当的反应，

就能更有效地与人类合作。另一个研究目标是实现多人互动的多台机器组成的体系“系统集合”。此外，人类与人工智能系统必须分享共同目标并且互相了解彼此以及当前状态的相关方面。要开发需要较少人类工程设计工作的系统，需要开展进一步的研究来汇总人类-人工智能系统的这些方面。

（二）开发用于增强人类能力的人工智能技术

以往的人工智能研究重点放在匹配或超越人类去执行狭义任务能力的算法。但为了开发增强人类多方面能力的系统，需要开展更多的研究。人类能力增强研究包括在静止设备（比如计算机）、穿戴式设备（比如智能眼镜）、植入式设备（比如脑机接口）以及具体的用户环境（比如量身定制的操作室）上运行的算法。

（三）开发可视化与人工智能-人机接口技术

更好的可视化及用户接口可以帮助人类理解各种来源的海量数据与信息。可视化及用户接口必须以人类可理解的方式清晰展示愈加复杂的数据及从中获取的信息。

（四）开发更有效的语言处理系统等

要达到人与人交流的水平，在语言处理研究方面面临着巨大的挑战。由于数据型机器学习方法的使用，实现了在安静的环境中实时识别流利的英语语音。但这只是向长期目标迈出的第一

步。除了解决嘈杂环境、地方口音、儿童语音等挑战外，还需要开发能够与人类进行实时对话的语言处理系统，以及以语言处理系统可存取的形式获得结构化领域知识来开展更多研究。

如要实现人机更直观自然的互动，还需要在许多其它领域取得进步。要建立适用于语音与文字的各种模式的计算模型，为判断情绪状态、影响和立场提供证据，并确定谈话与文本中所隐含的信息。最后，由于网络沟通方式与现实对话交流截然不同，因此必须完善这些环境中所使用的语言模式，从而使社交人工智能系统可以更有效地与人类互动。

战略三：了解和应对人工智能带来的伦理、法律和社会影响

我们希望自主型人工智能能够了解并遵守道德、法律与社会原则，其行为符合人类法律和道德规范。与其它任何技术一样，人工智能的使用必须要应对如何将法律与道德的约束应用于这项新技术的挑战。为了解人工智能系统的研发和使用对社会的影响，需要对该领域的关键信息技术开展进一步的研究。

（一）提高公平性、透明度以及设计问责制度

许多人对数据密集型人工智能算法的错误与误用以及性别、年龄、种族或经济阶层的划分表示质疑。在这方面，人工智能系统对数据的合理采集与使用成为了一项重要的挑战。但除了纯粹

与数据相关的问题之外，人们还纷纷质疑人工智能的设计能否公平、公正、透明并且有责任心。研究人员必须懂得如何设计这些系统才能使它们的行为与决策透明并且能够被人类轻易理解，从而可以检查其中是否含有偏见，而不是仅仅学习和重复这些偏见。这些严肃的智能问题牵扯到如何表达和“编码”价值理念与信仰系统。科学家还必须知道系统中的公正性与公平性在设计中能达到什么程度，以及如何在当前工程技术边界内实现这一点。

（二）建立符合伦理道德的人工智能

除了对公平和公正的基本假设之外，人工智能的伦理道德问题一直以来也被广泛关注。伦理道德是一个哲学问题，而人工智能依靠并且受限于工程技术。因此，在技术可行的范围内，研究人员必须努力开发符合或遵守现有法律、社会准则与伦理道德的算法和架构。道德准则的表述一般都十分模糊，并且难以转换成精确的系统和算法设计。当人工智能系统，尤其是新类型的自动决策算法，在面对独立且相互冲突的价值体系所产生的道德困境时，情况也会变得十分复杂。道德问题取决于文化、宗教和信仰，但为了解释和证明结论与行动的合理性，可以设计可接受的道德参考框架来指导人工智能系统的推理与决策。需要通过多学科方法生成体现合适价值体系的训练数据集，包括应对道德困境或冲突价值理念的首选行为示例。这些示例可能包含法律或道德方面

的“罕见个案”，并且贴上了用户可见的结果或审判标签。人工智能需要具备足够的方法应对价值理念冲突。为此，系统所包含的原则应可以应对严格规定无法解决的实际复杂情况。

（三）设计符合道德的人工智能架构。

为了确定如何设计出包含道德推理的最佳人工智能系统架构，必须在基础研究领域取得更大的进展。目前已提出多种方案建议，比如双层监控架构，即将运营型人工智能与负责运营行为道德与法律评估的监控代理分离。另一种首选方法是通过精确的人工智能主体概念框架保证人工智能行为对人类安全且无害的安全工程设计。第三种方法是使用理论集构建道德框架，结合对人工智能的行为进行逻辑约束，使之符合伦理准则。随着人工智能系统的通用性越来越高，其架构可能包含在多个审判层面处理道德问题的子系统，包括：快速响应模式匹配规则、用于行为描述和正当化的慢速慎重推理、表示用户可信度的社交讯号以及运营时间跨度更长的社交流程，从而使该系统能够符合文化规范。研究人员需要重视如何创造出符合道德、法律和社交目标的最佳人工智能系统整体设计。

战略四：确保人工智能系统的安全性

创建稳定、可靠和可信的人工智能系统，需要在系统被广泛

使用前研究确保其运行的安全性、可控性。与其他复杂系统一样，人工智能系统面临着设施安全与网络安全的重大挑战，其原因包括复杂且不确定的环境、难以预测的紧急行为、人机目标设定偏差、人机交互的影响等。为了解决所有问题，需要通过增加投资来提高人工智能的设施安全与网络安全。

（一）提高人工智能的可解释性与透明度

对于用户而言，许多算法（包括基于深度学习的算法）十分模糊，并且几乎没有任何现有机制解释结果。决策树归纳法等人工智能技术可以提供内置解释，但其准确度一般较低。因此，研究人员必须开发透明且可向用户解释结果的系统。

（二）建立信任

受到用户广泛信任的复杂系统往往比较透明、可审查、稳定可靠可恢复。为了取得用户信任，人工智能系统的设计师需要创建界面友好、信息丰富、准确可靠的系统。但操作人员为了了解系统运行和性能极限必须花时间开展充分的训练。

人类与人工智能系统之间的联系随着技术进步不断加强，所面临的信任挑战也不断增多，如与时俱进的能力、对技术应用的预见性，以及为了在研究设计、构建与使用方面达到最佳操作规范而制定管理原则和政策，包括为确保安全运营而进行合理的操作人员培训。

（三）强化认证与验证

人工智能系统需要新的认证与验证方法。“认证”确定系统是否符合正式标准，而“验证”确定系统是否符合用户的运营需求。安全的人工智能系统可能需要新的评估方法、诊断方法以及维修方法。此类系统可能需要具备自我评估、自我诊断和自我维修的能力，才能稳健而可靠。

（四）抵御攻击的网络安全性

嵌入到关键系统的人工智能必须具有稳健性才能处理事故，而且还应该能够抵御各种故意的网络攻击。网络安全工程需要了解系统漏洞以及潜在攻击者的行动。

（五）实现长期安全与价值定位

最终，人工智能系统可能会大幅修改软件实现“自我改进循环”。为了保证自我修改系统的安全性，需要进行更多的研究：检查系统行为是否与人类设计师的最初目标相一致的自我监控架构；防止系统在评估时发布的禁闭战略；系统推断用户价值理念、目标或意图的价值学习；以及能抵制自我修改的价值框架。

战略五：开发人工智能训练与测试共享公共数据集和测试环境平台

开发并提供用于人工智能训练与测试的资源，才能让我们持

续受益于人工智能。训练数据集和其他资源的种类、深度、质量与精确性会对人工智能的性能产生显著的影响。各种人工智能技术都需要高质量的数据资源和动态交互试验平台及模拟环境。这不仅仅是一个技术问题，还是一项重要的“公共利益”挑战。如果人工智能训练与测试被限制在寥寥几家已拥有宝贵数据集和资源的公司，而我们必须尊重商业和个人对数据的权利与利益，那么这项技术的发展就将举步维艰。如要开发适用于各种人工智能应用的优质数据集与环境，并且实现负责任的优质数据集以及测试和训练资源访问，就需要对此开展研究。此外，还需要通过更多开源软件库与工具包加快人工智能研发的进展。关键领域包括：开发并提供丰富的数据集以满足各种人工智能目标领域与应用领域的需求、使训练与测试资源符合商业与公共利益、开发开源软件库与工具包。

为了有助于为这一领域的持续创新提供支持，美国政府可以增加在开放式人工智能技术开发、支持和使用方面的工作。使用标准化或开放格式与开放标准表述语义信息的开放式资源尤其有用，包括领域知识本体。

政府还可以通过加快政府内部开放式人工智能技术的使用，鼓励进一步采用开放式人工智能资源，从而确保创新人员的入门门槛较低。政府应在可行时将算法与软件投入到开源项目中。由

于政府会产生具体的担忧，比如更加需要重视数据隐私与安全等，因此政府需要开发出缓解政府采纳人工智能系统压力的机制。比如，可以建立一支工作团队对政府各机构进行“水平扫描”，从而搜索部门内部的特定人工智能应用并且确定这些机构是否需要在解决具体的问题后才能采纳此类技术。

战略六：建立标准和基准评估人工智能技术

标准、基准、试验台以及人工智能社群对它们的采纳是指导和推动人工智能技术研发过程中所必需的。

（一）开发广泛的人工智能标准

为了匹配人工智能日新月异的能力及不断扩展的应用领域，必须加快标准制定。标准所提供的要求、规范、指导或特征，可以用于确保人工智能技术符合关键的功能性、互操作性、安全稳定性目标。标准的采纳能够增加人们对技术进步的信任并且加快互操作市场的扩展。所有人工智能的子领域需要制定更多人工智能标准。包括：（1）软件工程：管理系统复杂性、持续性、安全性，并且监视和控制紧急行为；（2）性能：保证精确性、可靠性、稳健性、可访问性与可扩展性；（3）指标：量化影响性能的因素以及标准的符合程度；（4）设施安全：评估系统的风险管理和危害分析、人机交互、控制系统以及合规性；（5）实

用性：保证接口与控制的有效性、高效性与直观性；（6）互操作性：定义通过标准与兼容接口可互操作的组件、数据和事务模型；（7）网络安全：信息的保密性、完整性和可用性以及网络的安全性；（8）隐私：控制信息在处理、传送或保存时的保护措施；（9）可追溯性：提供事件记录（其执行、测试与完成）与数据监护（10）领域：定义具体到领域的标准词汇和对应框架。

（二）建立人工智能技术基准

由测试与评估组成的基准为制定标准和评估标准符合性提供量化措施。基准通过推动应对部分战略性情景的技术发展促进创新，提供追踪人工智能科学技术发展的目标数据。作为与人工智能相关的成功基准之一，美国国家标准技术研究所（NIST）开发了一套综合全面的测试方法与相关性能指标，用于评估紧急响应机器人的关键能力。尽管这些工作为人工智能基准的建立打下了坚实的基础，但它们仍限制在具体领域的范围内。为了保证人工智能解决方案的广泛适用性与普及性，需要制定覆盖更多领域的标准、试验台与基准。

（三）增加人工智能试验台的可用性

《未来网络实验报告》指出了试验台的重要性：“研究人员必须通过试验台才能使用实际运行数据在现实世界的系统上设计和开展实验……并且在良好的试验环境中设计和运行情景。”

所有人工智能领域都需要有充足的试验台。政府拥有大量政府特有的任务敏感型数据，但大部分无法分享给外部研究社群。并为学术和行业研究人员建立合适的项目，使他们能够在特定机构所建立的安全、可控的试验台环境中开展研究。研究社群可通过访问这些试验环境共享和验证人工智能模型和实验方法，为人工智能领域科学家、工程师和学生提供独一无二的研究机会。

（四）让人工智能社群共同参与标准与基准的制定

推动标准化进程并且在政府、学术界与行业广泛推广使用需要政府的领导与协调。由于各政府机构根据各自的职责与使命以不同的方式与社群合作互动，因此可以通过协调利用社群互动加强其影响力。还需要通过此类协调综合采集用户提出的要求，预测开发者提出的标准并且普及教育机会。

行业和学术界是新兴人工智能技术的主要来源，因此鼓励并协调他们参与标准与基准制定十分关键。必须注意的是标准的制定与采纳以及基准活动的参与都需要付出成本。更新所有政府机构的采购流程，从而在征求计划书中加入具体的人工智能标准要求，这将鼓励社群进一步参与标准的制定与采纳。

战略七：更好的把握国家人工智能研发人才需求

人工智能研发技术进步需要有充足的研发人才。具有强大人

工智能研发实力的国家将主导未来的自动化技术，并在算法创建和开发、能力展示以及商业化等能力上领先一筹。近期商业与学术界多份报告指出，人工智能专家已供不应求，缺口越来越大。国家应该对人工智能研发人才的现状、供需力量、当前及未来的需求（包括学术界、政府和行业）展开研究。同时还需要了解人工智能研发人才的输送线，包括教育途径、再培训机会及多元化的研究。一旦能够更好地了解当前与未来人工智能人才需求，就能明确采取怎样的计划与措施来应对现有或预期的人员需求挑战。

要深入研究国家人工智能研发人才的培养与维系版图。尽管一些报告指出人工智能研发专家的缺口正越来越大，但目前没有官方数据表明人工智能研发人才、人才输送渠道以及人才供需力量的现状。考虑到人工智能研发人才战略规划中所扮演的角色，需要更好地掌握并维系健康的人工智能研发人员团队，这就要求联邦政府相应机构应采取措施来确保国家人工智能研发人员的健康培养与维系。

译自：*THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN, October 2016 by NSTC*

赛迪智库
面向政府 服务决策

研究，还是研究
才使我们见微知著

信息化研究中心	工业化研究中心	规划研究所
电子信息产业研究所	工业经济研究所	产业政策研究所
软件产业研究所	工业科技研究所	军民结合研究所
网络空间研究所	装备工业研究所	中小企业研究所
无线电管理研究所	消费品工业研究所	政策法规研究所
互联网研究所	原材料工业研究所	世界工业研究所
集成电路研究所	工业节能与环保研究所	安全产业研究所

编 辑 部：赛迪工业和信息化研究院
通讯地址：北京市海淀区万寿路27号院8号楼12层
邮政编码：100846
联 系 人：刘 颖　董 凯
联系电话：010-68200552 13701304215
　　010-68207922 18701325686
传　　真：0086-10-68209616
网　　址：www.ccidwise.com
电子邮件：liuying@ccidthinktank.com

报：部领导

送：部机关各司局，各地方工业和信息化主管部门，
相关部门及研究单位，相关行业协会

编 辑 部：工业和信息化部赛迪研究院

通讯地址：北京市海淀区紫竹院路 66 号赛迪大厦 15 层国际合作处

邮政编码：100048

联 系 人：蒯佳佳

联系电话：（010）88559594　18201126359

传　　真：（010）88558833

网　　址：www.ccidgroup.com

电子邮件：kjj@ccidgroup.com