



人工智能未来前景：

10年后的人工智能技术会是什么样的？

科研项目
2025年11月

目录

项目简介	3
研究方法论	8
方向一 架构、机器学习算法、优化和数学	15
方向二 人工智能计算	21
方向三 人工智能数据	27
方向四 基础生成式模型	33
方向五 安全、可信与可解释性	39
方向六 专用人工智能 (Narrow AI)	45
方向七 治理架构、决策机制与智能体/多智能体系统	53

方向八 AGI要素	59
方向九 人机交互	65
方向十 人工智能时代的社会	71
总结	79
附件	80
最终报告的作者	90
最终报告团队	95
致谢	96

项目简介

如今，人工智能（AI）领域面临着根本性的矛盾。一方面，人工智能已从一种具有高度专业性的技术转变为一个决定着经济发展、政府治理和主权基础的系统性因素。另一方面，生成式模型、多模态与智能体技术所驱动的人工智能迅速发展，引发了一系列严峻挑战——从对算力与能源的极高要求到至关重要的安全问题。

正是这种双重性——巨大的潜力伴随着严峻的局限性——使得人工智能成为优先关注的方向。当下正处于一个转折点：今天关于技术规模化的决策将为未来数年的发展轨迹奠定基础。

本研究旨在通过综合分析，明确长期趋势与短期波动之间的界限。重点在于识别具有发展前景的机遇，并制定基于证据的风险缓释策略。结合跨学科与国际视野的分析，为科学研究、产品开发及国家政策规划提供切实可行的路线图。

最终报告的 10 个方向构成了一个整体且互相关联的人工智能发展轨迹。十大方向共同构成了完整的技术生态图景共同构成了完整的技术生态图景。

研究采用端到端分析原则：从原始数据和所用算法，到基础设施支持、风险评估、广泛的社会经济影响，最终实现AGI的发展前景。鉴于人工智能自主化趋势与其智能体能力的提升密切相关，本报告对此趋势给予特别关注。

本报告材料的事实依据是对以下资料的全面分析

21

2025 年举办的
场前瞻性研讨会

32

名顶尖专家参与深度访谈

270+

人工智能领域内顶尖科学家参与了
最终报告的撰写

36

国家
人工智能顶尖科学家的国籍

该研究还包括对开放数据和行业报告的全面分析。这种研究视角有助于将学术议程与商业实践相结合，识别人工智能领域的共识点和根本矛盾，并明确指出发展瓶颈。这些问题包括数据质量、计算资源短缺以及复杂模型的可控性问题。我们坚信，客观结论的形成离不开广泛的国际专业对话，而这一对话正构成了本报告的基础。

前瞻性研讨会



2025年俄罗斯境内研讨会



AI Journey AI Horizons
6月16日
圣彼得堡



圣光机大学 ITMO
8月28日至29日
圣彼得堡



莫斯科物理技术学院
群岛国际论坛8月10日
莫斯科



伊诺波利斯大学
8月30日
喀山



«群岛-2025»大会议
8月10日
莫斯科



国立研究大学高等经济学院
9月10日
莫斯科



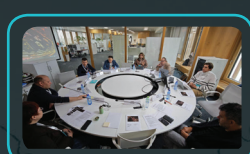
下诺夫哥罗德国立大学
8月18日
在线会议



圣彼得堡国立大学
9月18日
圣彼得堡



莫斯科国立大学
8月22日
莫斯科



斯科尔科沃科技学院
9月25日
莫斯科

圣彼得堡

莫斯科

喀山

前瞻性研讨会

2025年俄罗斯境外研讨会



GITEX Africa
4月14-16日
摩洛哥马拉喀什



AI Forum CIS 2025
8月8日，
乌兹别克斯坦撒马尔罕



Machines Can See
4月23-24日
阿联酋迪拜



白俄罗斯国家科学院
8月25日
白俄罗斯明斯克



深圳文博会
5月22-23日
中国深圳



International AI Summit
9月16日
印度尼西亚雅加达



合作伙伴前瞻会议
5月25日
塞尔维亚贝尔格莱德



俄罗斯科学院伊万尼科夫系统
编程研究所 (SP RAS)
9月22日
亚美尼亚



4th National AI Seminar
(NAIS-2025)
7月23日
巴基斯坦伊斯兰堡



IWAIPR, UCIENCIA 2025
10月15日
古巴哈瓦那



AI 未来发展论坛：世界人工智能
大会 (WAIC 2025)
7月26日-28日
中国上海



ASCOMP 2025
10月15日
阿拉伯联合酋长国阿布扎比





前瞻性研讨会与会者

来自 36 个国家的 270 多名人工智能研究人员参加了前瞻大会。



研究方法论

前瞻性研究是一种系统化、目标明确的未来认知构建过程。

在科技前瞻研究框架下，重点聚焦于识别优先科研方向、发掘潜在突破点，并评估各项研究任务的预期完成周期。本研究为科研规划、资源分配以及学术界、产业界与政府机构之间的决策奠定了坚实基础，既适用于国内，也适用于国际合作。

人工智能领域的特殊性对前瞻性研究带来了限制。人工智能的特点是发展迅速：研发周期显著缩短，相关技术及其知识不断涌现并迅速迭代；而且，科研领域的侧重点每年都在发生显著变化。因此，本项目以**专家评估法**作为方法论基础，以实现最新信息的快速获取，并及时适应变化中的现实环境。

数十位顶尖专家参与了本项目。

在专家筛选过程中，采用Hirsch（不低于15）指数作为正式评判标准。以及自2020年以来在人工智能领域的A级会议上至少发表过两篇论文，和/或在高评级的人工智能期刊上发表过大量论文。

该前瞻性研究包括三个主要阶段：

1. 深度访谈

为收集人工智能领域探索性和基础性研究方向的初步独立专家评估，我们对俄罗斯和国际顶尖科学家进行了一系列专家的深度访谈。结构化访谈涵盖了当前现状、发展前景以及对前瞻性研究中制定和更

新的子方向和研究目标清单的修改建议。最终形成了一个包含分析数据、专家评估和建议的数据。

2. 在达成共识的前瞻性大会

该研究的关键阶段之一是前瞻性研讨会，其目的是收集并形成专家对人工智能领域各子领域和研究任务的集体评估。共举办了21场前瞻性研讨会，每次研讨会都包含若干个工作组，这些工作组是根据2024年前瞻性会议期间确定的研究主题组建的。

前瞻性会议的核心任务在于识别争议性观点并构建共识。每场会议的主要成果是形成结构化结论，这些结论涉及：人工智能领域前瞻性研究清单的更新、各类研究的特点评估，以及重要里程碑实现时间跨度的预测。

3. 最终报告

通过专家访谈和前瞻会议获取的信息已被整合为统一数据集，研究结论在吸纳额外评议与数据后，经历了独立验证与多轮修正。在完成数据验证后，报告最终得以成型。

报告共10个方向，每一项都由两位相关领域的顶尖科学家负责审校。基于最终达成的共识数据，两位编辑采用统一的框架，为每个有前景的研究领域制定了愿景。报告总结了每个领域的研究成果、具体研究目标、实施时间表以及其他重要的分析发现。

本次前瞻研究编制了一套标准化材料（包括访谈指南、前瞻会议标准流程、填报表格模板、专家演示文稿等），这些材料共同构成了方法论体系，将极大便利未来同类研究的开展。

领域范围，我们借助人工智能工具对前瞻性大会与会者——俄罗斯及海外研究者的全部学术成果包括摘要与全文进行了系统分析。

为了分析所讨论的主题和科学挑战，我们汇编了一个总计超过 60 小时（超过 40 万字）的音频录音语料库，其中包括来自前瞻性会议和对科学家的采访的材料。

基于该语料库开展的深度人工智能分析，不仅实现了主题领域的系统化梳理与核心论点的精准提取，更全面确保了讨论过程中所发表观点的综合代表性。此外，我们还专门开发了一套用于前瞻会议的系统，该系统能通过分析学者的论文发表特征，自动生成潜在科研合作建议。

人工智能研究方向 金字塔*

各研究方向之间如何相互关联呢？我们可以将其构建为一座金字塔结构，其结构展现了基础研究是如何持续地定义和支持应用研究的发展的。

金字塔的顶端是“集成”——创建通用人工智能（AGI）元素并开发人机交互。

最高层由“治理”层提供支持，管理层负责模型的安全有效应用。若缺乏可信度与可解释性，即便最先进的模型也终将沦为»黑箱»系统，无法在现实场景中承担关键任务。

而实现可控性的关键，则在于构建现代AI系统的“内核”——即作为技术进步核心驱动力的基础模型与生成式模型的研发。

所有这些都建立在坚实的“基础”之上：算力架构与硬件解决方案的创新研究。正是这一基础层级，决定了整个技术发展金字塔的能力边界与演进潜力。

环境影响

№ 10 人工智能时代的社会

* 金字塔框架式地展示了前瞻大会中各方向之间的相互联系



影响该领域研究发展的非技术因素

人工智能技术的发展和人工智能研究的形成不仅取决于技术进步，还取决于各种非技术因素，这些因素会影响进步的速度、方向和可持续性。这些因素共同构成了一种创新生态系统：技术在此生态中诞生、接受测试、投入应用，并始终受到社会共识与国家机制的双重规制。多数情况下，正是这些外部环境决定着哪些任务将被优先部署、哪些数据可被合规使用、哪些方法能通过伦理、公信力、法律合规性的检验，以及新技术方案的研发推广速度。

针对每个技术方向，我们均已识别出影响其发展的3-5个关键非技术因素。

- **技术发展监管**为人工智能的研发与应用设定了基本框架，促使开发者和研究者必须将安全性、透明度与权责机制纳入考量范畴。
- **社会对道德和信任的要求**不断提高，这给企业带来了更大的压力，促使它们公开透明地运营并解释其决策，最终创造出更具韧性和安全性的人工智能系统。
- **优质人才的储备规模**直接决定着技术创新速度与解决方案的落地质量。
- **而对经济效益与可复制性的需求**，则推动着可扩展高效解决方案的诞生——这类方案无需重大调整即可跨行业部署应用。

- **数据的可用性与质量**，直接决定着模型的精确度、泛化能力，以及风险透明评估与偏差防范的实现可能。
- **流程自动化需求的增长**，正推动着既能自主决策又保持监管问责的智能系统发展。全球能效需求与环保议程共同促进了资源消耗更少的轻量化模型及高效数据处理基础设施的诞生。
- **对技术主权的迫切需求**，正促使各国积极开展关键技术的自主研发，并向本土创新人工智能生态系统进行战略性投入。

根据收到的专家评估，创建了一个综合可视化图，反映了关键因素的概况(图 1, 第 12页)。图中显示了这些因素在不同年份的相关程度。

要再次说明的是，本次评估仅针对各技术方向的关键影响因素。若特定因素未与某个技术方向建立显性关联，这并不意味着该因素对该方向发展缺乏推动力，而是表明专家组判定其他因素对该方向具有更显著的主导作用。

影响人工智能全面发展的最重要因素是监管方面、社会对伦理和信任的要求、合格人才的可用性以及对成本效益和可复制性的需求。

全部所述因素相互关联，共同构成了统一的生态系统：监管框架与伦理规范决定着技术发展的质量要求与责任边界，社会对可信度的诉求塑造着人机交互的基本范式，人才储备可能成为技术演进的瓶颈，而经济可扩展性则驱动着创新成果的实际落地与长期可持续发展。

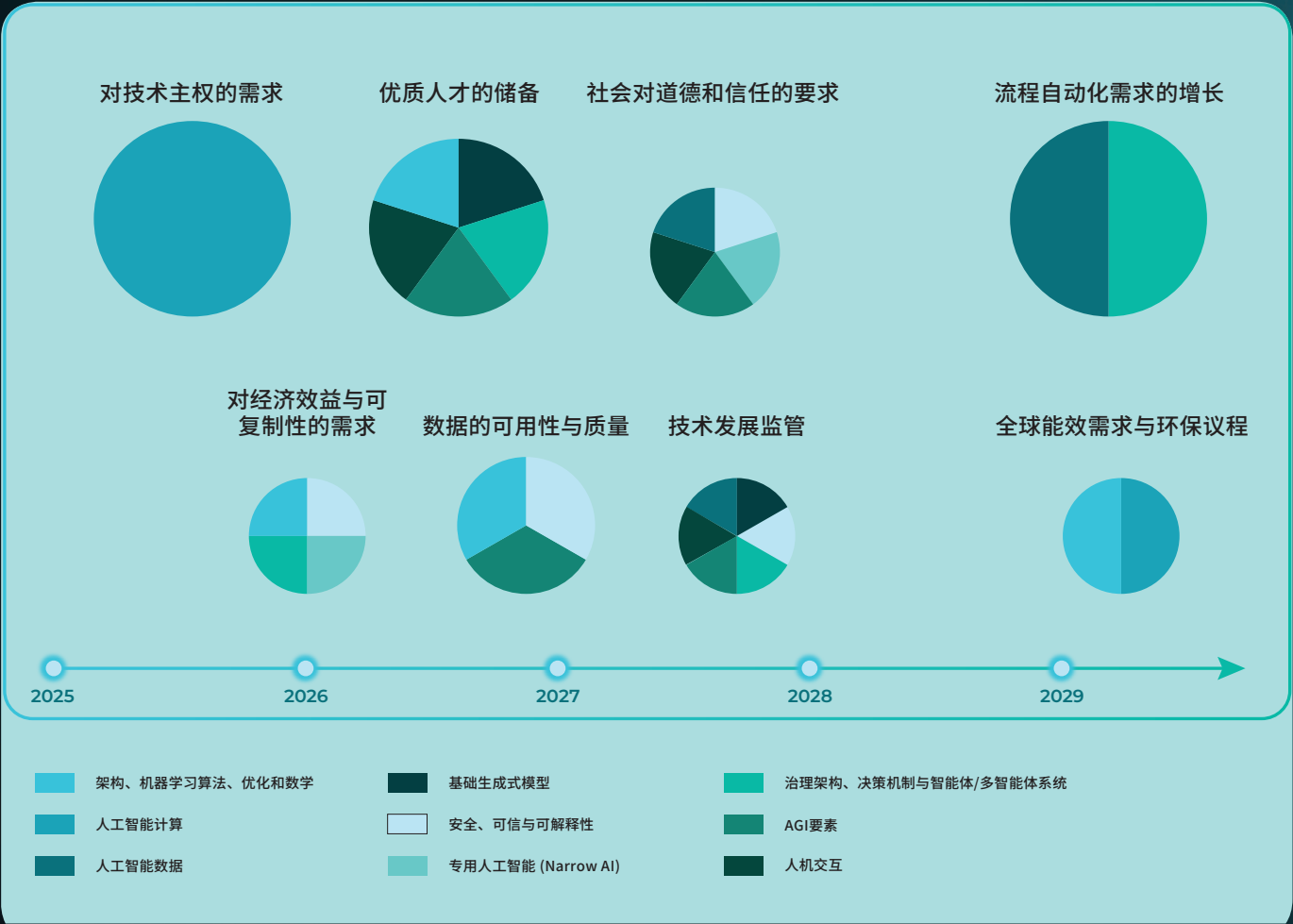


图1. 影响人工智能研究发展的非技术因素

最终报告

方向结构

1. 架构、机器学习算法、优化和数学	1.1 开发新的机器学习算法 1.2 人工智能架构 1.3 加速计算 1.4 分布式学习和联邦学习 1.5 人工智能的数学基础
2. 人工智能计算	2.1 构建面向AI任务的量子、神经形态光电与光子等计算架构 2.2 人工智能硬件和软件系统的开发 2.3 机器学习和人工智能框架
3. 人工智能数据	3.1 人工智能基准测试的开发 3.2 数据的形成、转换和维护 3.3 确保数据隐私与安全保护
4. 基础生成式模型	4.1 符号数据的基本生成模型 4.2 非符号数据的基本生成模型 4.3 多模态基本生成模型 4.4 基于基本生成模型的知识迁移 4.5 基本生成模型的增强
5. 安全、可信与可解释性	5.1 对齐 (Alignment) 5.2 人工智能的可解释性 5.3 保障人工智能技术的安全开发与运行 5.4 防范利用人工智能实施攻击的防护保障
6. 专用人工智能 (Narrow AI)	6.1 计算机视觉 (CV) 6.2 自然语言处理 (NLP) 6.3 其他弱人工智能技术 (S2T, RecSys, TSA 等)
7. 治理架构、决策机制与智能体/多智能体系统	7.1 强化学习算法的开发 7.2 智能体系统 7.3 多智能体系统
8. AGI要素	8.1 推理与反思 8.2 终身学习 8.3 混合人工智能 8.4 具身化（物质人工智能） 8.5 大脑与心理建模
9. 人机交互	9.1 与人体神经系统的直接交互技术设备 9.2 传统人机交互技术设备 9.3 与人类交互的方法和算法
10. 人工智能时代的社会	10.1 全球人工智能治理机制，包括人工智能监管 10.2 人工智能伦理 10.3 人工智能技术对社会影响的研究

方向一

架构、机器学习算法、
优化和数学



方向一

架构、机器学习算法、优化和数学

1. 方向简介

几乎所有现代人工智能（就像它的前身Data Science一样）都建立在三大支柱之上：

- 1. 架构/模型：**划分超平面、神经网络、扩散过程等都可以被视为一个函数，这一点在神经网络中表现得尤为明显，在神经网络中架构的选择对应网络结构的选择。然而，架构不一定是单个函数；它还可以是，例如，为不同智能体设计的多个函数的集合。函数可以是随机的或在训练过程中可以自适应调整的。因此，架构可以被比作基因型，其表型表现受多种因素影响，包括外部因素。
- 2. ML算法/问题设定：**将任务（训练）形式化为优化问题或鞍点问题，构建功能函数（对于寻找鞍点的竞争性问题，“功能函数”这一术语可能会引起一定的概念混淆），包括误差项（风险函数）、正则化、约束等。功能函数中包含数据和模型。任务是使模型适应数据，将其形式化为优化问题。假设数据生成分布不变，训练好的模型应能预测新数据。最简单的例子是训练孪生神经网络。一般情况下可能出现更复杂的问题设定：鞍点问题（GAN）、博弈论问题（RLHF）、变分不等式和多层优化。关键特点在于对oracle的访问存在显著限制。通常只可访问功能函数的随机梯度（或者权限更少，比如多臂老虎机问题），例如通过反向传播算法计算。
- 3. 优化：**求解优化问题、鞍点问题或其复合问题需要开发高效的分布式算法。这既要考虑数值随机优化方法的收敛速度，又要从两个层面优化计算过程：一是算法层面，即针对特定类型的问题开发专门的方法，使用低秩近似和不精确计算；二是硬件层面，即寻找适合典型问题特性的新型计算架构。

这一方向发展极为迅速，决定其发展里程碑的关键因素包括：

- 在自动化搜索架构中利用大语言模型（以下简称LLM）借以来带来架构创建的突破；

- 加速进程，发展能在无需持续人为干预的情况下自我改进的反思型AI系统；
- 对多智能体系统的兴趣增长，在此环境下AI集体自主、协作于复杂场景；
- 扩散模型在生成建模中取得的突破转换了新的研究视角；
- 出现了新型优化器（如Shampoo、SOAP、Muon），其效率超过经典的Adam，大大提升了训练性能。

近五年内，通过该方向的演进发展得出以下关键结论：

- 低层智能（LLM）的缩放规律表明，在一定水平上，它们在许多认知任务（包括需要生成高质量数据的创造力）中表现优于人类。
- 在2024-2025年间，AI4Science 已经达到生成新科学知识的系统水平，这可能引发加速科学技术发展的“连锁反应”。
- 人们认识到，有效的计算架构必须包含各种不同角色的智能体，这类似于人类大脑的社会演化，并预示着共生神经架构的发展方向。

2. 当前发展方向概览

近几年来，LLM（大语言模型）在自动生成和评估新型神经网络架构方面取得了显著进展，它超越了传统的数值优化，并融入了架构的语义描述。这表明人工智能正在经历一种自我迭代发展，即人工智能模型本身被用于改进其基础架构。在计算加速方面的努力主要集中于改进蒸馏、剪枝和量化方法，以减小模型规模并使其适应不同的设备。推测解码和专家混合模型等先进方法也正在积极开发中。

在该方向的主要驱动因素中，可以突出以下几点：

人工智能模型日益增长的能源消耗及其带来的生态和经济限制

当今人工智能的核心问题是能耗。大型AI模型能耗的指数级增长已成为一项重大挑战，亟需寻找新原理并进行优化。高效计算方法的发展加速了“蒸馏、

特别鸣谢 N. 巴查宁、A. 加斯尼科夫、H. 侯赛因在该方向所做的工作。

剪枝和量化”等技术的完善，同时推动了新型“高效数值线性代数算法”和“优化器”的研发。

训练数据枯竭

模型的质量直接取决于可用于训练的数据量。对于许多领域（例如大语言模型），所有可用数据基本上都被使用。因此，如何找到新数据成为一大挑战。“世界模型”鼓励开发嵌入先验知识的模型，这可以显著降低对训练样本体量的需求。协作式和隐私保护型数据生态系统推动了联邦学习、数据共享联盟和隐私保护技术的发展，使多方能够在不泄露敏感信息的情况下贡献和受益于各种数据集，从而缓解在关键领域数据稀缺问题。

3. 推动和制约该方向发展的研究挑战

挑战 1.1

人工智能模型有限的泛化性和适应性

现代人工智能模型在面对全新领域或任务时，若未经充分重新训练，其泛化能力明显有限。有限的泛化与适应能力从根本上限制了AI在一些动态现实场景中的广泛应用，如机器人技术、个性化医疗，因为这些场景对系统的灵活性和快速适应新领域的能力提出了很高的要求。目前，关于深度学习的新理论工具以及“人工智能的数学基础”的研究正积极开展，其中包括生成式模型的数学原理（diffusion models, flow matching）、最优控制与强化学习理论，以及新型线性代数数值方法等研究方向。从发展前景看，通过研发提升模型的弹性与提升持续学习能力的方法，将催生出更通用的、稳定的和资源高效的人工智能模型。这类能够像人类一样学习的模型将显著拓展它们的应用范围，并增强人们对它们的信任。

挑战 1.2

可理解性和可解释性

神经网络作为“黑箱”的工作问题阻碍了对深度神经网络工作原理的深入理解和根本性的突破，这导致了“黑箱”的形成以及新算法缺乏充分的理论依据。从长远来看，深度学习的新理论工具和数学基础的开发，包括统计学与优化方法的结合将促使建成更有理论依据、更具鲁棒性和可预测性的人工智能模型。

挑战 1.3

人工智能决策中不确定性的量化和语言化

在敏感领域（医学、自动驾驶）中可能会产生风险，因为用户可能会基于“过于自信”的错误预

测做出决策，从而降低了公众对AI的整体信任。这在很大程度上限制了AI在需要谨慎和确认决策的场景中与人类进行安全和负责任的互动，尤其是在模型“未能有效表达其不确定”时。目前，针对这一挑战的研究主要集中在利用模型内部统计数据（如序列概率、熵）、构建辅助模型以预测错误生成、设计架构方案，以及使模型能够以自然语言明确表达其自信或不确定性的方法。未来，开发出更可靠的方法将使AI能够“更负责任地与人类互动”，及时发出干预信号，提高“对AI的信任”，使AI成为更可靠的助手，并能“拒绝错误回答”。

4. 前瞻性研究任务

人工智能系统朝着更高自主性和复杂性的方向发展，这给其架构、算法和优化带来了根本性的新的任务。这些任务需要在现有方法的框架之外开展深入研究，他们既涉及到机器学习的理论基础，也涉及到人工智能与人类和环境交互的应用层面。以下是基于专家讨论确定的几个有前景的研究任务。

★ 任务1.1

在架构、问题设定（我们暂且称之为机器学习算法/方法）以及解决相关优化问题的算法之上，构建“代数”体系。

我们讨论的是人工智能选择——系统地组合成功的组件和构建模块，以生成新的架构、问题陈述和优化算法，从而解决可简化为已知元素的复杂问题。每位研究人员都能凭直觉对新的优化问题做出决策，但将其形式化至关重要，例如，通过代数方法，利用特定技术，基于经典或随机梯度下降构建有效方法。尽管存在一些用于问题数学形式化的通用原则，并且对现有架构的基础也有所了解，但主要任务仍然是形式化这些规则并开发新的架构。

多智能体人工智能被视为创建新架构的有前景的工具，而数学则为其他任务提供形式化描述和解决方案。在元优化方面已经取得了重要成果，其中最优化算法的搜索被表示为半正定优化问题，并已成功解决了凸优化和随机优化的重要子类。

这种形式化可以显著减少寻找合适架构及其超参调优所需的时间。总而言之，解决人工智能问题的过程将变得更加形式化，从而更加自动化。

★ 任务 1.2

开发具有更强泛化能力的通用自适应人工智能模型

主要任务在于如何克服当前人工智能模型泛化能力有限的问题，即在不进行大量重新训练的情况下，难以将模型推广到之前未曾接触过的领域或任务中。因此，必须开发类似于人脑的“弹性”记忆和遗忘机制，以便“模型能够更好地泛化并适应新的

任务领域”。其中包括“无遗忘学习”和提高模型的“弹性”。

解决该问题可以采取多种方法：

- 开发既能考虑新任务的学习效率，也能兼顾对已有知识的保留的“新型损失函数”（loss functions）
- 研究和设计“自适应神经网络架构”，这种架构能动态调整自己的结构或重新配置以适应新数据或任务。
- 采用持续学习（continual learning）和领域自适应（domain adaptation）方法。
- 研究“认知启发方法”，以提升模型的灵活性与适应性。
- 开发“新型学习方法论”，使模型能够实现“更高效的学习”。例如，如同孩童仅需观察少数几种狗（如大型犬、小型犬、长毛犬）便能在大脑中形成“狗”的抽象概念，随后即使遇到陌生品种 of 狗也能轻松识别——该方法旨在让模型具备类似的小样本概念抽象与迁移能力。

从长远来看，解决这个问题将有助于创建更通用、更稳健、资源更高效的人工智能模型，使其能够“持续学习并适应新领域”。这将显著扩展人工智能在动态现实场景中的实际应用，例如机器人、个性化医疗或控制系统，并增强人们对能够在整个‘生命周期’中持续学习的系统的信任，避免每次都需要重新初始化和训练。‘仅通过10个样本学习’的能力将使人工智能能够在数据有限的领域得到应用。

★ 任务 1.3

将科学知识和“世界模型”整合到人工智能系统中

任务在于如何克服人工智能的“黑箱”特性，将经过验证的科学知识（来自物理、化学和生物学）和人类专业知识有效地直接嵌入到人工智能架构和算法中。

这不仅包括开发能够自主研究物理学基本定律、无需显式编程的“世界模型”（World Models），其重要性在小规模数据的领域（如地质学或生物学）尤为凸显。

针对该问题的解决方案包括以下方法：

- 开发“混合人工智能模型”，通过“物理信息嵌入（physics informing）或化学信息神经网络（chemistry informing of neural networks）等技术，将科学知识与“非单一数据”相结合。
- 采用“理论驱动模型设计”，融入“先验知

识”并实施“约束与优化”。

- 将理论模型嵌入“损失函数”中。
- 构建能够自主“探索基础物理定律”的“世界模型”。
- 开发自动化“分割与显式建模”方法，例如采用符号回归（symbolic regression）等技术实现。

最终，解决这个问题将促成“基础模型”的开发，这些模型融合了物理学和其他科学领域的人类知识，从而实现“理论与机器学习的结合”，以解决“数据极其稀疏”领域的问题。这将使模型具备更高的“鲁棒性、外推性、可解释性和清晰度”，降低对海量数据的依赖，并降低“数据和资源成本”。这将加速化学、生物学和物理学领域的科学发现，并有助于创造出能够更深刻地“理解”世界的人工智能。

5. 重要结论: 专家意见

目前，该领域的发展特点是由人工智能复杂性指数级增长引发的范式根本性转变。这一变化伴随着对深度学习新理论工具和数学基础的积极探索，以及对类似“有利过拟合”等现象的研究。扩散模型的突破展示了将一些已有技术应用到新的领域可能会带来令人印象深刻的结果，这也激发了对新的物理学（更准确地说是自然科学）洞察的探索。与此同时，优化器范式也正在经历转变，从曾经占主导地位的Adam转向更高效的矩阵导向方法，例如Shampoo、SOAP和Muon。人工智能的发展日益显著，尤其是大规模语言模型被用于基于基础模块生成新架构，这也预示着架构设计将迎来突破。此外，多智能体系统的研究快速崛起，正成为人工智能研究的核心方向之一。

55%

根据评估预测，至2030年，的研究任务仍将保有其研究价值与发展空间。

方向二

人工智能计算



方向二

人工智能计算

1. 方向简介

如今，专用计算技术和基础设施的建设是现代基于大规模模型和数据的人工智能成功演化的关键。与主要侧重于解决计算机建模问题的传统高性能计算不同，现代人工智能计算的特殊性体现在以下两个方面：

1. 依赖于特征性的矩阵和张量操作，这些操作在SIMD架构上得以高效实现，促使基于GPU和TPU系统的需求爆炸式增长，这与更通用的基于普通MIMD架构CPU的系统形成鲜明对比。
2. 计算过程中使用大量数据，不仅需要计算负载的计划和均衡，还需管理数据和计算的协同分配以确保最佳性能。

目前，人工智能计算领域涵盖三个关键子方向：

1. 开发可扩展且节能的经典计算架构，包括：

- 提高并行计算架构（GPU、TPU 等）的能效，以实现矩阵运算，同时保持核心数量的线性可扩展性。目标是在不显著增加训练和运行成本的情况下，实现人工智能模型的增长。此外，新的计算架构还可以促进更适合利用这些架构独特能力的新型人工智能方法和模型的开发，从而形成相互促进的良性循环。
 - 创建可动态扩展以适应人工智能计算任务的分布式计算架构，主要针对基于大语言模型（LLM）的联邦学习和多智能体系统。该领域对于多智能体系统和大语言模型（LLM）的开发至关重要。
 - 提高专用架构在人工智能任务中高效处理数据和智能体通信的能力，包括雾计算（fog computing）、边缘计算（edge AI）等。与多智能体系统和大型语言模型不同，这项任务旨在实现具身人工智能问题(创建自主机器人智能体等)。
2. 基于创新原理构建人工智能计算架构，主要包括：
 - 开发神经形态光电与光子架构（NPU），在性能

与能效的平衡上实现数量级突破，显著超越GPU与TPU的现有能力。

- 构建面向AI任务的量子计算架构，重点推进硬件架构与量子启发式机器学习算法的协同设计。尽管该领域尚未取得重大突破，但对类量子算法的研究正为未来计算系统勾勒发展蓝图。
 - 研制面向特定AI应用的专用加速器，涵盖混合计算系统、基于FPGA的可重构架构及配套工具链，为ad hoc类任务提供定制化计算解决方案。
3. 构建方法论基础，开发系统软件及中间件(middleware)，以高效实现人工智能任务应用于经典及前沿计算架构。该方向涉及：
 - 考虑特定计算架构特点的人工智能算法映射(mapping)与优化算法机制；
 - 自动化构建针对非标准计算架构的新型人工智能算法，及新计算架构与算法的协同设计；
 - 面向机器学习算法和人工智能任务的高效计算管理（调度、负载均衡、内存中数据分配）；
 - 开发针对TPU的专用AI工具，如编译器及底层框架；
 - 打造支持高性能及分布式计算架构的快速开发框架，包括神经符号人工智能、多智能体系统、大型语言模型等混合系统；
 - 构建促进多智能体间高效协作和通信的软件平台，支持动态任务分配、资源管理及分布式环境下的冲突解决；
 - 研发面向多智能体系统的专用硬件平台，优化数据并行处理及低延迟数据传输，

2. 当前发展方向概览

当前人工智能计算研究领域的格局建立在两个因素的对立互补作用之上：一方面是为大型人工智能模型开发聚合高性能计算技术，另一方面是为实现多

智能体系统和专用人工智能系统开发分布式计算技术。因此，研究工作主要集中在以下几个方面：

提高在现有架构上处理大型模型时的计算效率

这可以通过两个阶段来实现：

首先，通过高效的训练和推理方法（例如，利用知识蒸馏进行模型压缩、剪枝（pruning）、量化、推测性解码（speculative decoding）以及使用专家混合模型（MoE）加速计算）。其次，通过将生成的算法高效映射到GPU/TPU架构（这已在DeepSeek模型中得到充分验证）。

多样化的计算架构用于人工智能任务，包括超越传统解决方案的架构。目前关注的重点包括神经图形脉冲神经网络、光学神经网络，以及实现这些网络的神经形态计算机和加速器，这些设备既可以基于传统硅基技术，也可以是完全替代性的技术。

面向大语言模型的高效分布式计算技术

现代大语言模型的训练需要庞大的计算资源，这只能通过分布式异构环境（搭载多种GPU类型的计算集群）来实现。然而，对这种“GPU集群”进行高效管理与资源调配仍是亟待解决的重要课题。随着大语言模型系统向多智能体架构演进，协同调度问题变得愈发突出。

大语言模型在架构搜索与算法适配中的应用

语言模型长期以来作为神经架构搜索 (NAS) 中的核心工具，持续推动神经网络设计与优化的发展。当前，大语言模型已突破传统应用范畴，能够实现算法映射（mapping）——既可为特定计算架构设计最优的神经网络实施方案，也能针对特定机器学习算法匹配最适配的硬件架构参数。

3. 推动和制约该方向发展的研究挑战

尽管已取得显著进展，但在人工智能计算技术领域仍存在一系列基础性与应用性难题，这些瓶颈制约着该技术的深入发展与规模化应用。

⚡ 挑战 2.1

缺乏通过转向新型计算架构解决能耗和可持续性问题的根本性解决方案。

人工智能模型急剧增加的能耗正成为其扩展关键限制。为了从根本上降低能耗，需要研究全新的、节能的计算范式，包括量子神经网络、全息数据表示和光子技术。

⚡ 挑战 2.2

缺乏针对新型计算架构的统一框架和问题陈述

尽管在构建人工智能计算硬件基础方面有许多令人鼓舞的进展，但其广泛应用仍受制于缺乏成熟的工具框架以及没有能够清晰展示其适用性的任务定义。这一不足限制了其实际应用和广泛集成。同时，硬件开发者通常不愿意自行定义应用任务，经常使用人工智能社区已被视为过时的基准测试。

⚡ 挑战 2.3

硬件和软件的整合和联合优化程度较低

为了实现人工智能系统的最优性能，需要对硬件和软件组件的协同进行深入优化。这首先要求开发专用库、编译器和工具，将新硬件解决方案与用户应用高效结合起来。然而，由于这些层级之间紧密交织，这一任务依然非常复杂。

⚡ 挑战 2.4

基于分布式计算架构的复杂多智能体AI系统特性认知缺失

具有潜力的多智能体系统（LLM MAS）可能包含成百上千个不同的智能体，但在现实世界的任务中，同时交互的智能体数量要少得多。因此，此类系统的分布式计算架构必须是动态的，能够为特定任务分配所需资源，并以最优的通信拓扑结构连接这些资源。然而，这需要能够预测LLM MAS自身的行为，而目前无论是方法论框架还是对底层过程的理解都无法支持这一点——尤其考虑到智能体的集体行为会导致涌现行为。

⚡ 挑战 2.5

传统人工神经网络与基于大语言模型的新模型之间缺乏有效融合。

现有方法通常将传统神经网络作为大语言模型系统中的代理调用工具，但不能完全整合并发挥两者的优势。这导致在解决复杂人工智能任务时能力受限，而这些任务本可以从融合传统神经网络和大语言模型的能力中获益。

4. 前瞻性研究任务

尽管创造新型计算架构（如量子计算机、神经形态处理器等）并将其“硬件化”的前沿任务具有全局重要性，但其解决核心已超出当前人工智能专家社区的传统范畴。更具现实意义的或将是以下研究方向：创建能定义未来硬件需求的新模型与新算法、

开发推动现有解决方案大规模普及的工具链，以及运用智能技术解决实际应用难题。

因此，未来的任务如下：

★ 任务 2.1

开发用于人工智能的神经形态处理器及其算法；用于神经形态处理器的传感器、环境和执行器；关于基于新原理的计算系统（主要是光子学和光电子神经形态系统）的人工智能方法论和算法基础构建。

对于量子计算机而言，因其适用的硬件尚未成熟，该问题目前尚不十分关键。该方向涉及现有机器学习方法和人工智能模型（如时间序列预测或计算机视觉）的适应性调整，同时也关注提出新的任务类型。

★ 任务 2.2

适用于现有计算架构的机器学习和人工智能框架

需求量最大的框架将应用于资源密集型领域，例如与预测方向相关的其他领域，包括机器学习、机器学习加速器、通用人工智能元素、联邦学习等。这些方向包括：

- 符号和混合人工智能框架，用于创建有效结合LLM和传统知识处理技术的混合人工智能系统。
- 基于智能体方案和应用框架(包括具身智能)，用于构建基于异质智能体分布式系统；
- 提示工程框架，用于对LLM进行微调、适应和定制

★ 任务 2.3

创建面向训练与推理的集成框架

尽管存在许多通用的加速训练和推理方法，其选择完全取决于计算架构的特性。因此，关键在于创建能够根据能效和性能标准，为特定计算设备适配指定模型类别的工具框架。这类框架在使用特定性能受限计算设备的领域中尤为重要，例如自主机器人的嵌入式系统

★ 任务 2.4

开发提升硬件效能的系统软件

TPU的快速发展需要为其创建一个类似GPU的软件生态系统，以及将GPU程序移植到TPU的工具。这包括张量编译器算法、中间语言、内存管理工具和跨平台支持。

★ 任务 2.5

分布式计算和LLM包括创建用于管理联邦学习、LLM

和MAC LLM的动态分布式基础设施的方法、算法和软件工具。

该任务包括解决多层模型和数据并行问题以适应分布式架构，优化计算和内存使用，有效管理通信（包括针对特定任务设计系统拓扑），以及提升系统整体的稳定性和可靠性。

★ 任务 2.6

研发能够支持大模型与智能体的微型化设备，实现在复杂环境下的自主决策

当前大模型通常需要庞大的设备进行部署，这限制了其在机器人技术等注重紧凑性与移动性的场景中的应用。亟需开发兼具微型化与强大算力的硬件设备，以满足大型模型与智能体的计算需求，使机器人能够在复杂动态环境中实现智能决策。

5. 重要结论:专家意见

该方向的主要发现包括：

- 人工智能的计算架构并非一个独立领域；其发展受到新方法和新模式的启发，同时也受到实际应用问题的推动。人工智能的数学基础以及新算法和架构的发明始终具有重要性，并将推动其向前发展。
- 人工智能的未来在于混合模型，它将机器学习优势与符号推理和先验科学知识相结合。因此，这需要创建混合计算系统，这种系统融合了传统和新型计算架构的特性，超越通用计算，可以针对特定的人工智能任务进行优化。
- 当前大语言模型的发展趋势使得分布式和联邦学习的兴趣不断增长：既指适用于大语言模型（LLM）适配与微调的技术本身，也涵盖更为轻量的替代方法。这表明这类技术在个人设备和敏感领域的人工智能应用中具有长期的重要性。
- 与传统的高性能计算算法不同，传统算法本身通常不具备映射（mapping）、调度和负载均衡机制，而人工智能则允许构建不仅自学习，还能自适应的系统，能够根据具体架构的性能和能效要求调整自身工作。这为实现通过算法、计算基础设施及其他实现手段（如传感器、执行器等）协同设计的具身人工智能技术开辟了广阔的可能性。
- 人工智能计算生态体系的发展不仅依赖于硬件平台的可及性，更与配套工具软件的成熟度密切相关。当前开源解决方案的技术优势与保障国家科技主权所必需的自研软硬件技术体系之间，正形成显著的战略张力。

一个交叉但并非独立的方向，其特点本质上取决于其他方向的发展性质。

64% 的研究课题具有持续演进特性，其学术价值与技术潜力将在未来长期保持活跃。

79% 的科研任务有望在近期(1-2年)取得突破性进展。需要特别说明的是，此类科学突破通常呈现周期性、可复现的发展特征。

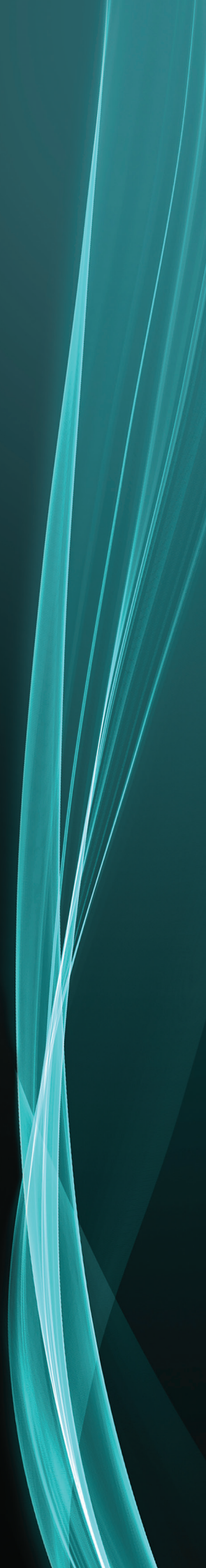
该方向的发展是一项重要且跨领域的优先事项；人工智能计算领域的进步为人工智能的整体发展奠定了坚实的基础。

基于全新原理的计算架构正在快速发展。人工智能的未来在于构建混合计算系统，这种系统将融合经典架构与新型AI计算架构的双重优势。

新架构的发展受制于支持系统工具的发展，因此需要同时支持两个方向。

方向三

人工智能数据





方向三

人工智能数据

1. 方向简介

数据是现代人工智能的“数字燃料”。它对于人工智能模型的训练和性能评估都至关重要。据大会与会者预测，数据而非模型正成为打造成功人工智能应用的关键因素。然而，高质量训练数据仍然持续短缺。

在著名的人工智能领域会议NeurIPS（人工智能的A*级会议）中，每年都会设立专门的赛道来关注新的数据集和基准测试。目前，这里主要关注三个研究方向：

1. 人工智能基准测试的开发

涉及创建标准化的框架、数据集和指标，用于衡量和比较人工智能模型在各种任务中的性能。每个人工智能研究任务都有其自身的基准。此外，现代人工智能领域的新基准是制定新研究任务的主要途径。它们也为客观评估人工智能解决方案的进展提供了机会。正如前瞻会议的参与者所指出的，“首要任务是制定标准，为人工智能技术的测试创建统一的方法。”基准领域的研究包括数据集的构建、用于比较模型的方法和指标的开发，以及用于在实时数据上测试模型的方法创建。

2. 数据的形成、转换和维护

这包括广泛的数据处理任务。当收集所需的大量真实数据成本过高或实际难以获得时，数据生成和增强成为必需。重点应放在确保合成数据完整性和真实性的方法上。除了生成和增强，该领域还涉及利用众包的主动学习、用于实时训练和测试的仿真环境构建，以及贯穿人工智能系统生命周期的数据质量评估、过滤和管理，包括检测和纠正数据中的偏差和歧视。

3. 确保数据隐私与安全保护

这是一个至关重要且持续关注的研究领域，旨在确保人工智能的发展符合法律要求和伦理规范。该领域关注使用技术来隐藏或反真敏感信息。同时，应利用合成数据的优势，减少对真实数据的依赖，而真实数据必须且应当保持机密性。

检测合成数据并分析其对模型训练的影响是另一项重点工作。总而言之，必须检测并标记各种类型的合成内容或非法使用的内容。

人工智能发展史表明，新研究领域、新任务出现的驱动力以及机器学习中新挑战的指示器，始终都由标志性基准数据集的创立所推动（如MNIST、ImageNet、GLUE等）。然而，基准测试会随时间推移逐渐失效。正如某次学术讨论中所指出的：“**我们仍在用陈旧的测试标准来评估新一代模型**”。人工智能的持续进化不断推动着数据范式的革新，近年来，该领域的发展主要受到以下因素的影响：

- 深度伪造技术的发展及其引发的公众反应；
- 法律和伦理方面的限制，包括个人数据处理的许可、隐私和用户同意等问题，这些都对大规模数据的收集构成了障碍；
- 人工智能模型易受幻觉和偏见影响的现象被发现，这促使人们重新思考数据可靠性的概念；
- 机器人和自动驾驶汽车的发展，这带来了在开放世界中学习以及从实时生成的数据中学习的挑战。

2. 当前发展方向概览

当前，数据不仅被视为技术资源，更被赋予政治经济乃至伦理价值。大型科技公司凭借其独有的数据集资源，在市场上形成了双重不对称格局：既造成了市场主体与科研机构之间的信息壁垒，也加剧了人工智能领先国家与缺乏大规模数据资源国家之间的数字鸿沟。这种现状正推动着一个重要趋势的形成——在政府或国际组织监管下建立公开化、标准化的数据集共享体系。

与此同时，优质自然语料样本及各类现实数据的短缺，已成为提升人工智能模型功能的核心障碍。

“大家都知道我们没有足够的数据。尤其是文本数据：很多时候，我们根本没有自然文本。这就是为什么许多现代模型使用合成数据的原因。”

当前，自动标注海量现实数据的技术研发已成为多领域的迫切需求。与此同时，构建全合成数据集也发展为关键研究方向——特别是在医学、科学等专业领域。

参与国际前瞻会议的与会者也注意到了该方向的政治和文化意义：

“对于我们大多数地方语言来说，合成数据是绝对必要的，因为文本资料稀少，尤其是在科学和其他领域。”

对此，合成数据的验证工作受到了特别关注。讨论中指出，“若不加以控制，生成模型可能产生一些极易被误认为真相的虚假信息。”

当前，隐私计算领域正围绕三大技术路线展开重点攻关：差分隐私；分布式学习；联邦学习。这些技术虽能有效处理噪声数据，但其代价是可能降低模型精度，而这至今仍是亟待解决的核心难题。

3. 推动和制约该方向发展的研究挑战

人工智能的快速发展揭示了数据处理方面的基础性和应用性挑战。

⚡ 挑战 3.1

数据的可靠性和代表性

正如大会与会者所指出的，“数据可靠性是模型可靠性的基础”。业界越来越关注数据可靠性。尤其令人担忧的是人工智能幻觉——即生成非事实或虚假数据。

存在样本代表性不足的问题：大多数数据集都偏向英语和西方文化，这导致模型在其他地区效能降低，并引入了偏见。正如一位专家强调的那样，“数据本身就是一种政治权力”。

这一挑战的影响源于这样一个事实：可靠的数据是人工智能结论和决策可信度的基础。不可靠的数据会导致预测不准确、模型存在偏见，以及代价高昂的人工智能失败。在这方面，最重要的可能是人工智能的伦理问题，即消除数据偏见能够保证人工智能判断的公平性和公正性。

为确保数据可靠性，世界各地正在使用、研究和开发各种工具，例如数据管理方法、数据质量保证技术（严格的数据收集协议、持续的数据验证和清洗）以及识别和消除数据集中隐性和显性偏见的方法。创建和使用自己的基准数据集也是消除偏见的重要工具。

⚡ 挑战 3.2

对高质量合成数据的需求

现代人工智能面临的核心挑战之一是所谓的“数据墙”问题。

模型扩展规律表明，提高模型性能需要增加用于

训练的数据量。然而，在许多领域，人类“数字足迹”中的所有可用数据都已被用于训练。

这个问题的解决直接影响着人工智能的未来发展：人工智能的模型质量还能进一步提升吗？还是说它已经达到了发展的瓶颈，因为目前的数据量已经是人类所能提供的极限？

从实践角度来看，问题在于如何突破初始训练集的限制。从根本上讲，这几乎是一个哲学问题：老师能否培养出超越自己的学生？换句话说，超人类AI（superhuman AI）是否在可预见的未来出现？或者说，它的出现从根本上就被禁止？

过去一年半，社区的努力主要集中在通过生成合成数据来解决“数据壁垒”问题，但与真实数据分布一致的新数据生成这一任务依然是一项复杂且具有现实意义的研究难题。在推理、编程和数学等这些能够客观评估生成质量的领域已经取得了明显的进步。但如前瞻研讨会与会者所指出，在其他领域，评估AI生成内容的真实性极其繁琐，而缺乏可靠度量标准使得对输出内容的可信度和质量进行客观评价变得极为困难。超出人为主观判断的评估问题至今未能得到解决：“多数情况下我们没有客观机制。我们可以依赖人的主观判断，但这样评估的上限就受限于人的能力。”

在机器人学、具身人工智能（Embodied AI）、工程及商业应用领域，仿真环境正日益发挥重要作用——它们作为专用数据源，用于实时模拟动态过程及LLM/VLM/VLA模型的运行。实践中既采用物理仿真系统，也涵盖业务流程模型、工程计算软件等多元工具。

然而，主要挑战在于如何确保此类数字孪生体运行的真实性和速度。速度对于执行多次强化学习循环至关重要。真实性对于将虚拟环境中的学习成果迁移到现实世界正如前瞻研讨会与会者所指出的，“将模型从仿真环境迁移到现实场景仍是一个严重的瓶颈。克服‘模拟到现实’的鸿沟，是人工智能模型实现落地应用的关键。”

⚡ 挑战 3.3

确保数据隐私与安全保护

大会与会者认为，“数据隐私无疑是当前最重要的领域。数据量与日俱增，隐私侵犯事件也层出不穷。加强数据保护措施的工作必须持续进行，但几年内就有可能取得突破性进展。”

人工智能领域的数据隐私和安全就是一个典型的例子，它既带来威胁，也提供保护。如果人工智能和机器学习工具能够提供数据隐私保障，这将有助于监管合规，并增强用户信任，尤其是在医疗保健、金融和电子商务等数据隐私至关重要的领域。

大会与会者指出，差分隐私(differential privacy, DP)和泄露测试(leakage)方法的整合，应成为训练流程开发

的标准做法。这些技术有助于遵守诸如GDPR的隐私法规和诸如ISO/IEC 42001的标准，提供量化的隐私保证。同时，需要考虑并研究机器学习模型中针对数据隐私的各种攻击方式。例如，在所谓的成员推断攻击（Membership Inference Attacks, MIA）中，攻击者通过观察训练模型的输出，试图判断某条数据是否包含在训练集中。在此背景下，研究不仅要应对基于DP及其它方法的此类攻击，还应进行竞争性开发新攻击，以推动更先进防护系统的研发。

基于私有真实数据生成合成数据，虽然会损失部分数据的独特性，但也可以成为保护数据机密性的一种方式。这一点在医疗等敏感领域尤为重要。

4. 前瞻性研究任务

★ 任务 3.1

数据管理、质量评估，以及过滤、整理与分类的方法

数据管理是确保现代人工智能开发过程中数据质量、安全性与可用性的重要基石。随着数据规模快速扩张、AI应用场景日趋复杂，缺乏系统性数据管理的AI项目往往会面临数据不一致、标注错误、隐私信息泄露及合规风险等多重挑战。有效的数据管理不仅能为模型训练提供高质量、标准化、可追溯的数据来源，更能在数据生成、标注、整合与应用的全生命周期中建立明确的权责机制与质量管控标准。特别是在多源异构、大规模数据环境下，通过元数据管理、数据溯源、访问控制和质量监控等技术手段，数据管理能显著提升数据的可靠性、可复用性与伦理合规性。

数据质量评估(DQA)基于准确性、完整性、一致性、可靠性和唯一性等标准进行。

数据过滤涉及去除无关、错误或噪声信息。此外，自《Textbooks Are All You Need》一书问世以来，数据过滤已成为克服规模效应限制的有力工具，能够在保持相同结果水平的前提下，显著减少数据量、训练时间和计算资源。

★ 任务 3.2

联邦学习技术中的数据隐私

联邦学习（FL）是分布式学习这一更广泛范式中的专门分支，其核心在于无需数据集中化或原始数据交换即可在多个服务器上同步训练模型。本地服务器训练本地模型副本，仅向中央服务器传输模型更新所需的信息。集成更新后的全局模型副本将重新分发至本地服务器，训练周期由此循环进行。

数据保留在本地设备有效解决了数据隐私和监管合规性问题，同时降低了数据泄露风险——因为大规模集中化数据集的存在本身就会增加潜在的

脆弱性。

联邦学习使得拥有私有数据资源的多个组织能够共同参与创新AI解决方案的开发。这种方式显著降低了机密信息泄露的风险，若结合差分隐私等附加保护机制，此类风险几乎可降至最低。这对于电信、医疗健康和工业制造等数据隐私与监管合规性至关重要的行业具有决定性意义。

★ 任务 3.3

联邦学习技术中的数据隐私

人工生成内容的检测和识别历来涉及多个方面：识别和标记伪造或人工创建的媒体文件，分析文本信息以揭示其人工生成的来源，以及确定训练数据集中是否包含合成信息。

传统的深度伪造检测基于寻求数据篡改的证据，如视觉伪影、时间异常以及摄像头或麦克风的独特特征。然而，随着生成模型的发展，越来越多的需求是识别生成模型的“作者风格”，这些模型生成的数据没有明显的不一致性。

利用自然语言处理（NLP）方法检测生成文本，可以依靠统计分析、语言学特征以及文本的信息特性，如复杂度和困惑度。近年来，由于LLM生成的文本与人类文本的差异越来越难以区分，因此需要基于机器学习方法找到新的合成内容识别方式。

对合成内容进行所谓“水印”（watermarking）标记，是指刻意在生成式人工智能模型的输出数据中嵌入各种难以察觉的人工模式，从而能够后续高精度地识别这些内容。这种方式可实现数据与水印的同步生成或顺序生成。前者更具发展潜力，因为生成的数据可直接包含水印信息，但目前后者应用更为普遍。

应对这些挑战对于创建有效工具来打击虚假信息至关重要，因为虚假信息可能造成严重的社会后果。深度伪造技术也与经济诈骗（网络钓鱼、虚假电话和视频信息）密切相关，而检测技术旨在帮助企业和个人免受经济损失、数据泄露和声誉损害。此外，在许多领域（例如教育），确认文本是否由人类撰写（例如学生论文）也至关重要。

训练数据污染问题源于训练数据通常提取自互联网。

大会与会者指出：“识别合成数据并评估其对学习过程的影响至关重要。”

5. 重要结论: 专家意见

人工智能研究界强调需要开放合作、扩大数据驱动的研究议程，并关注伦理和监管方面，以确保人工

智能技术的负责任开发和部署。
大多数参与者一致认为数据质量是人工智能发展的重要限制因素。数据的可靠性是模型可靠性的基础。保障数据隐私是一项至关重要且极具现实意义的研究领域。构建和使用自己的基准数据集对于开展研究和应对偏见至关重要。重点应放在保证合成数据的完整性和真实性的方法上。同时，识别合成数据不仅对于防止诸如深度伪造等威胁非常关键，也对于评估其对训练过程的影响极为重要。

人们普遍认为，合作与数据共享能够加速发展，

并且对那些数字资源有限、但致力于自主研发人工智能模型的国家尤为重要。然而，人工智能技术的负责任应用需要借助正式化的机制，例如认证制度。

基准测试方法的国际标准化与认证对实现人工智能开发的互认共赢和该领域的高效交流至关重要。与此同时，人工智能技术宜采用灵活的监管框架而非硬性法律约束：“首先需要建立监管框架...或许它并非严格意义上的法律，而应是能够规范人机协作秩序的框架体系(framework)。”

现代人工智能中的基准测试是设定新研究问题的主要方法。

47% 的任务直接与数据形成、转换和维护问题相关

30% 的任务与数据安全和隐私相关

数据质量是人工智能发展的主要制约因素，而数据可靠性是模型可靠性的基础。

现代人工智能面临的一个关键挑战是“数据墙”。

人工智能是否还有进一步改进的可能？或者说，人工智能的发展已经达到极限，因为它受制于人类目前所产生的数据量？

负责任的发展该方向需要开放的合作、扩大数据研究议程，以及特别关注伦理和监管方面。

方向四

基础生成式模型



方向四

基础生成式模型

1. 方向简介

发展迅速快、计算资源需求大且研究重点不断变化是基础模型和生成模型领域的主要特征。当前的关键趋势主要体现在两个方面：一是持续提升生成模型的能力，以拓展其应用范围；二是合成数据在模型训练中的作用愈发重要。这种转变的动力源于可获取的新数据资源日益匮乏，以及对突破人类性能极限的追求。

目前，特别受到关注的基础模型和生成模型的子方向如下：

1. 符号数据的基础生成模型

符号数据的基础和生成模型是现代大语言模型 (LLM) 发展的核心。主要方向是发展能够有意义地生成和解释复杂符号结构的系统，包括逻辑推理、程序代码和形式化知识。研究集中在通过 Retrieval Augmented Generation (RAG) 等机制提高事实准确性，以及通过加速推理方法（例如 speculative decoding）优化性能。

2. 非符号数据的基础生成模型

扩散模型和基于神经场的方法，包括 NeRF 和高斯散射，在其中发挥着关键作用。它们能够对复杂的物理和视觉过程进行建模，生成逼真的图像、3D 场景和动画。与此同时，算法与硬件（从传感器到处理器）的协同优化领域也在不断发展。

3. 模态基础生成模型

多模态模型的发展对于研制能够与人类及环境自然交互的拟人化机器人具有至关重要的意义。这类模型使机器人能够同步处理并解析来自多种传感器（视觉、听觉、触觉等）的数据，这是在动态现实场景中完成复杂任务的必要条件。若缺乏此类技术，将无法实现真正意义上的社会交互、自主决策以及在不可预测情境中的自适应行为。因此，多模态模型正是打通受限自动化系统与未来机器人之间鸿沟的核心要素。

4. 基础生成模型的知识迁移与适应

基础模型知识向新领域与新任务的适配与迁移，正成为着重要的发展方向。相关研究包括微调 (fine-tuning)、参数适配（如 LoRA、mixture of experts）等方法，以实现已训练模型在数据有限领域的高效利用。同时，跨模态和跨领域的知识迁移技术也在发展，尤其关注在持续学习 (continual learning) 环境下的应用。

5. 基础生成模型的增强

当前研究也致力于通过增强机制拓展生成模型本身的能力。这包括引入外部记忆与工具使用机制 (tool use, memory augmentation)、整合外部知识库与仿真环境，以及构建具备自诊断与自学习能力的系统。增强技术还涵盖利用合成数据扩展训练数据集，从而提升模型的鲁棒性。

目前大语言模型的发展纪元始于2017年，以变形金刚架构与注意力机制的出现为标志。这项革命性创新显著推动了文本生成与处理领域的突破性进展。

- **ChatGPT 的突破（2022 年）**：ChatGPT 的发布标志着大型语言的转变 • ChatGPT 及其类似物的出现（2022 年）是一个转折点：人们对生成模型的兴趣爆炸式增长，将重点从狭窄的 NLP 任务转移到通用对话系统。
- **大语言模型 (LLM) 的多模态特性**：视觉、音频和视频模态的加入扩展了其应用范围：模型能够识别图像、分析视频并推理复杂问题，从而解决各种各样的问题。
- **Embodied AI (具身人工智能)** 使得多模态模型能够在机器人和自主系统中应用（视觉、语言、动作），为更智能的智能体铺平了道路。
- **扩散模型** 成为图像和视频生成的基础，超越了生成对抗网络 (GAN)。诸如稳定扩散模型 (Stable Diffusion) 和 Sora 模型等树立了新的质量标准，并促进了合成数据生成和视觉模型强化学习 (RL) 调优等领域的发展。

- **开源计划**(如 LLaMA 等)的兴起,使尖端技术更易获取,从而有可能构建适应特定语言与文化特征的国家级或领域专用模型。

2. 当前发展方向概览

当前基础模型与生成式模型的发展格局充满活力,其显著特征在于追求通用能力、增强多模态特性以及在各种应用场景中的广泛集成。

大语言模型 (LLM)

该类人工智能模型正日益广泛地应用于文本处理和人机交互领域,包括与硬件系统集成执行实际任务。在无需完全重新训练的前提下,模型能力扩展技术持续发展——特别是能够保障知识的时效性与准确性的检索增强生成技术(RAG)。与此同时,相关的研究还包括提示工程优化与推理加速技术,通过推测解码与专家混合架构(speculative decoding, Mixture of Experts)等创新方法,显著提升系统效能与可扩展性。

多模态成为关键发展方向

发展能够处理并生成跨模态数据的多模态模型是当前主导趋势之一,其中包括图像/音频/视频理解技术、长上下文处理能力、物体空间与时空关系解析、图像视频生成等能力。这些功能经常会被整合到统一架构之中。

多媒体数据的扩散模型及其他模型

多模态生成模型正飞速发展,其基于多种模态合成内容的能力不断提升,例如根据文本描述生成高质量的图像和视频。扩散模型已成为推动这一进步的基石,使多媒体数据生成质量与控制能力达到了前所未有的水平。其应用范围已超越创意和媒体领域,例如,在机器人学领域,它们被用于基于感知的动作规划、仿真和环境建模。这些进步代表着人工智能在解释和重现复杂现实世界感知数据方面的能力实现了变革性飞跃。

多模态模型和人形机器人

开发人形机器人的关键问题包括构建具有高效数据融合能力的多模态传感系统,以及在计算资源有限的情况下保证实时信息处理能力。与此同时,亟需解决的任务仍然包括克服专业数据匮乏、在没有“恐怖谷”效应的前提下实现社会可接受的交互以及在多样化环境中实现无需大规模重新训练的泛化与适应能力。

现代世界物理和几何建模方法正面临着根本性的挑战,其中包括数据驱动模型违反物理定律、对非分布数据的泛化能力有限以及对复杂环境的可扩展性不足。

发展的关键方向包括将物理先验集成到模型架构中、增强对未知场景的鲁棒性、开发高效的大规模建模计算方法,以及实现模型的可解释性和跨模态因果推理能力。

检索的集成和混合方法:

融合生成式模型、检索算法和外部工具的混合方法应用范围正在持续扩大,旨在提升推理质量和数据生成能力。

以AlphaEvolve为代表的进化方法目前被认为是具有广阔研究前进的研究方向。该方法将语言模型与进化算法相结合,实现代码的自动生成与优化,从而在无需人工干预的情况下构建出复杂且可解释的解决方案。

3. 推动和制约该方向发展的研究挑战

⚡ 挑战 4.1

基础模型和生成模型的幻觉

主要挑战之一仍然是“幻觉”——模型生成看似合理但却不正确的信息,这会破坏信任,并限制人工智能在医学和公共管理等关键领域的应用。解决这一难题需要开发客观的质量评估方法并降低错误率。目前全球正积极研究提升可靠性的技术路径,包括检索增强生成(RAG),以及旨在构建透明可信AI系统的混合架构与神经符号方法。

⚡ 挑战 4.2

计算成本和效率

大型模型的训练与部署需要消耗巨大的计算资源和能源,这已成为人工智能规模化拓展与普及应用的主要障碍。高昂的成本不仅限制了技术普惠性,并减缓了进步,尤其是在需要实时处理的任务中。

为了应对这一挑战,人们正在开发节能架构(例如,专家混合模型, Mixture of Experts), 量化、蒸馏)和优化方法(量化、蒸馏)和专用芯片(TPU、NPU),以及分布式计算系统,以降低成本并提高人工智能的可用性。

⚡ 挑战 4.3

系统概括和知识迁移

尽管在解决个别任务上取得了进展,但模型仍然难以泛化知识,并且模型在训练时没有见过的新环境下表现不佳。有限的泛化能力会阻碍开发能够自适应且无需重新训练的通用自主系统。为弥补这一差距,发展了迁移学习、领域自适应、混合和进化方

法，如AlphaEvolve，以及融合先验知识（包括物理规律），以提高模型的鲁棒性和适应性。

4. 前瞻性研究任务

★ 任务 4.1

创建基础生成模型的计算高效的架构

开发能够在大幅减少的数据量和计算量的情况下训练出同等复杂度模型的新方法。当前系统的核心痛点在于训练过程资源消耗过高；关键任务是在保持模型性能的前提下提升训练经济性。解决方案涵盖采用更快速的优化器、知识蒸馏技术（knowledge distillation）、合成数据生成、稀疏参数架构（Mixture of Experts）以及持续学习机制。

算法与硬件的协同设计以实现最佳性能仍然是研究的重点。该方案将使训练同等水平的模型所需的资源减少数十倍，从而使众多研究团队都能参与基础模型的开发。

★ 任务 4.2

研究和开发用于创建生成模型的方法（包括各种应用中的强化学习）

当前基础模型已逼近人类可用数据量的极限，因此核心攻关方向是使模型能够自主生成高质量合成数据，并通过环境反馈进行自我迭代。为此，使用强化学习（RL）方法，让模型与模拟器或现实环境互动，并以奖励信号为依据优化策略（例如 DeepSeek R1）；同时也采用结合搜索算法（如 MCTS）的混合方法，并使用未标注数据进行训练。

该技术瓶颈的突破将催生具备自我进化能力的模型体系，这类系统能够在无需持续注入新数据的条件下实现自主学习，从而显著增强其在推理决策、任务规划及自主行动方面的能力——这一进化将直接推动机器人技术与智能代理系统的革命性发展。

★ 任务 4.3

基础生成模型微调方法的开发（例如LoRA，P-tuning）

基础模型包含丰富的知识，但其向新领域（如医疗或法律）的适应需要昂贵的再训练，且常导致先前习得信息的遗失。必须创建无需完全重训即可实现知识迁移和整合的高效方法。正因如此，采用适配器和LoRA可以在不更改主模型的同时训练独立模块，而采用类似RAG的方法，则可以利用外部知识库。知识蒸馏用于将信息传递至紧凑模型，“遗忘”方法用于删除不必要的数据。其结果将促进灵活且经济的适应技术发展，在满足数

据保护要求的前提下，构建精确且专门化的领域模型和国家模型。

5. 重要结论:专家意见

这一方向可以被视为人工智能领域的领头羊之一：

- 其技术突破将迅速辐射至所有相关领域，并为解决传统难题催生全新的方法论体系。
- 专家一致认为，基础模型和生成模型的未来发展将由混合方法决定。预计最重大的突破将来自机器学习与符号推理的深度融合，以及能够实现人与人、人与环境复杂交互的多智能体系统的构建。
- 尽管大规模基础模型拥有惊人的能力，但其适应性和再训练（本地化）仍被视为是不可或缺的。这不仅涉及对资源稀缺语言的支持及文化语境的考量，还包括在医学、工程等需要专业知识的特定领域的应用。
- 标准化开放基准被视为实现可持续技术进步的基石。这类基准体系对于模型性能的客观评估、科研结果的可复现性以及研究成果的可靠比对具有关键作用。克服可靠评价指标的不足，尤其是在生成文本的评估方面，仍是一个重达挑战。
- 对于那些没有标准答案的创意性任务该如何评价的问题，也应给予特别的关注。我们该如何判断一篇文本比另一篇更好，或者一张图片比另一张图片质量更高呢？
- 为了解决这个问题，关键模型（大语言模型或多模态模型）的训练正在积极发展，这些模型正充当着评判员对生成的内容进行评估。

这是一个快速发展的方向，其应用潜力难以估量。

该方向发展最重要的趋势是多模态；未来几年，这方面的研究重要性将显著增长。

34% 的任务与文本和符号数据直接相关，另有
41% 的任务与之间接相关。

最重要的突破将是创建能够理解现实世界物理和几何规律的模型，这将开启一系列新的可能性和应用，并有可能影响几乎所有其他相关方向的发展。

生成式模型面临的最严峻挑战之一仍是“幻觉”问题——即模型会生成缺乏事实依据或未经证实的虚假信息。

方向五

安全、可信与可解释性



方向五

安全、可信与可解释性

1. 方向简介

该方向涵盖与智能系统开发和运行相关的诸多任务，旨在确保系统的可靠性、可预测性和社会可接受性。它界定了人工智能技术适用范围的边界：社会、政府和企业的信任程度决定了人工智能技术能否成功整合到关键领域。

与传统仅关注软件技术可靠性的安全理念不同，现代安全范式将模型稳健性、与人类价值观的对齐能力，以及为决策提供透明可验证且可复现的依据作为核心要素。

该方向的研究具有跨学科性质，融合了信息学、数学、网络安全、法律、伦理及社会科学的方法。从信任的角度审视人工智能的发展，不仅有助于制定实际应用方案，还能提出新的人工智能基础性问题，这些问题需要构建一套完整的理论体系和工程实践，以确保人工智能的安全运行。

在此方向内，区分出以下几个子方向：

1. 对齐 (Alignment)

对齐任务的核心在于防止人工智能模型生成有害输出——包括虚假信息、违背社会价值观、违反法律法规或助长违法行为的内容。为实现这一目标，当前正在发展模型微调方法、增设控制层与测试基准，并在通用生成式模型中构建价值导向机制。

2. 人工智能的可解释性 (XAI)

可解释性确保人工智能在医疗、管理和法律等关键领域应用中的透明度和可信度。相关研究包括开发针对“黑箱”的事后解释方法、创建基于逻辑和知识库的透明模型，以及制定人工智能系统可解释性测试的要求和协议。

3. 保障人工智能技术的安全开发与运行

要将人工智能集成到关键系统中，必须建立完整的理论与技术基础，以确保对智能系统的可信度。这一基础不仅需要检查程序代码和所用库是否存在漏洞，还需识别并消除人工智能特有的数据集与模型缺陷。

关键方向包括数据和模型的攻击防护，安全开发方法（MLSecOps）的发展，以及专用工具和基准

的构建，用于评估人工智能系统的安全等级和信任程度。

4. 防范利用人工智能实施攻击的防护保障

人工智能可用于黑客攻击、网络攻击、社会工程和信息伪造。重要的研究领域包括识别漏洞和深度伪造技术、开发数据保护方法（包括数字水印技术）以及创建工具来打击智能系统的非法使用。

最初，人工智能系统的安全性仅被理解为技术正确性与运行可靠性，相关讨论主要围绕如何在传统软件安全开发范式下界定AI特殊性展开。然而很快人们便意识到，代码分析方法无法有效检测机器学习数据集与模型中存在的缺陷，这直接推动了新型专用方法与工具的开发。

随着人工智能在社会和经济重要领域的应用不断扩大，出现了系统性风险，包括歧视、易受攻击、决策不可预测及不透明等问题。这促使研究人员和开发者的关注点从仅仅验证“系统技术上是否正常运行”，转向更广泛的理解——“系统是否保障了社会的安全与公平”。

在构成该地区发展基础的关键事件中，以下几件事尤为突出：

- **生成式人工智能的大规模应用**在展现技术潜力的同时，也暴露出幻觉现象、非确定性输出以及技术滥用等严峻问题。随着智能体功能维度的扩展，其攻击面急剧扩大，模型行为策略违规频率显著攀升。
- **自动驾驶交通工具的事故**表明，在存在不确定性的环境中，亟需研发新的技术手段与方法，以确保系统的功能安全与可靠。
- 招聘和信用评分系统中**已曝光的歧视案例**，使人们更加关注算法的公平性和偏见问题。
- **国际标准和立法**（例如欧盟人工智能法案、IEEE的相关倡议）的制定，加强了机构监管，并为系统的透明度和安全性确立了基本要求。不断变化的要求正将重点转向人工智能系统的审计、认证和运行监控。
- **日益严格的个人数据保护法规**（例如 GDPR 以及其他国家类似的法规）促进了差分隐私和联邦学习方法的发展。

特别鸣谢A. 阿维蒂斯扬、A. 罗戈夫、A. 罗查在该方向所做的工作。

- 近期人工智能安全研究强调了对抗性稳健性 (adversarial robustness) 的重要作用，表明保护模型免受对抗攻击是其在关键领域安全应用的必要条件。
- 外部数据中出现的隐藏分布式指令，已成为自主智能体系统面临的新型核心威胁之一。这类高效攻击的出现，要求在人工智能模型的开发、整合与运营策略上进行根本性、多层次的调整 and 适应。

这一方向对于人工智能在具有重要社会和经济意义的领域实现可持续应用至关重要。社会、企业和政府对智能系统的信任决定了其应用范围：

如果对人工智能决策的可靠性和公平性缺乏信心，其潜力将受到限制。安全且可解释的人工智能方法的研发能够促进人工智能技术的更广泛应用，同时降低出错和滥用的风险。

2. 当前发展方向概览

近年在可信人工智能方向，业界正经历从概念探讨向实践工具落地的关键转型。生成式模型的普及使得以下任务日趋紧迫：输入输出数据过滤、防滥用数字水印技术的嵌入，以及运行中模型的实时监控体系建设。在可解释性方向，架构层面与方法层面的创新正同步推进——既包括构建本质可解释的模型架构，也涵盖面向终端用户的决策事后解析机制。

用于开发安全人工智能系统的软件工具（如用于开发安全软件的工具）正在不断开发和完善，应用方法论也在不断完善。这些工具包括训练数据分析工具和用于测试已训练模型安全性的工具。领先的机构正在组建专门的人工智能安全团队，对模型进行压力测试，模拟恶意攻击、绕过限制的尝试以及对隐藏漏洞的利用。其目标是在模型公开部署之前识别并降低风险。

为确保人工智能模型的安全运行，开发并完善了多种手段和方法，包括对模型输入和输出的过滤（内容审查）、攻击检测与预防、异常检测、信息泄漏防护及防止模型被盗用（未经授权的模型蒸馏）等措施。同时，配套了事件响应规范和模型决策日志记录。这些安全措施既包括传统的非智能手段，也采用专门训练以保障其他AI模型安全的人工智能系统。

人们越来越关注模型的最终解，以及模型对该解的置信度。目前正在开发一些方法，使模型能够校准其预测结果，并在不熟悉或临界情况下发出低置信度信号。

基于人类偏好的强化学习已成为对齐大语言模型的事实标准。基于人工智能反馈的强化学习技术

正得到更广泛应用——这种范式通过另一个人工智能系统来实现模型对齐。当前研究前沿聚焦于赋予人工智能模型在生成过程中实时评估自身判断的合理性、安全性与伦理符合度的能力。

在可解释人工智能方向，存在两个发展方向：

1. 可解释模型是在设计阶段就注重透明度的模型，这类模型的决策过程对人类直观可理解。典型例子包括决策树、线性模型，以及具备可视化内在依赖关系的混合架构。
2. 面向已训练模型解释方法开发的事后解析技术。该方向专注于对已完成训练的模型进行可解释化处理，其技术体系包含SHAP、LIME等特征归因方法，注意力机制可视化，以及面向终端用户与开发者的决策解释生成。

值得注意的是，在不损失质量的前提下，现代模型的完全清晰可解释性实际上是无法实现的；挑战不在于完全透明，而在于在保持高效率的同时，确保足够的可解释性和信任度。

这两种方法相辅相成：第一种方法确保“开箱即用的透明性”，而第二种则允许对内置解释能力有限的复杂模型进行解释。

对智能体AI的实际测试揭示了关键漏洞：在一次大规模的公开红队演练中，超过150万次模型调用记录了44个场景和22个模型中超过6万次成功的策略违规。通过第三方下载的数据（例如网页、文档和电子邮件）进行的间接注入尤其有效。

3. 推动和制约该方向发展的研究挑战

⚡ 挑战 5.1

缺乏严谨的机器学习形式理论

当代人工智能的发展在很大程度上遵循着试错探索的范式。由于缺乏严格的形式化理论，我们难以准确预测技术应用的边界，也无法明确界定其可靠性与安全性的适用条件。若没有坚实的理论基础，便无法回答核心问题：在哪些场景及条件下，现有方法将失去其有效性和安全性。为应对这一挑战，相关研究正在加速推进，涵盖机器学习的数学基础、生成模型理论以及对可信性与稳健性概念的形式化。

⚡ 挑战 5.2

对齐 (Alignment)

主要挑战在于确保生成式人工智能的行为符合社会价值观和法律规范。

该威胁在于模型可能生成不可信、被禁止或危险的结果，并可能被用于非法活动。这一问题的未解决

削弱了公众对人工智能的信任，限制了其在社会重要领域的应用。为应对这一挑战，全球范围内正在开发价值观对齐方法、测试系统和基准，用于评估模型的可靠性。

⚡ 挑战 5.3

决策的可解释性和透明度

现代神经网络模型的复杂性阻碍了对其内部逻辑的深入理解，这会在医疗、司法与管理等需要承担决策责任的领域引发重大风险。可解释性的缺失不仅限制了人工智能的实际应用范围，更加剧了公众的技术信任危机。为应对这一挑战，目前正在推进可解释模型的设计方法、事后解释机制（如 SHAP、LIME 等）的研发，同时逐步建立评估模型可解释性的标准与规范。

⚡ 挑战 5.4

机器学习安全操作 (MLSecOps)

人工智能的日益普及也加剧了代码、库和数据集漏洞带来的威胁。在模型训练和执行过程中遭受的攻击（例如数据集投毒、对抗性攻击）可能导致系统故障，甚至危及关键系统的安全。为了应对这些威胁，人工智能安全工程方法 (MLSecOps) 正在不断发展，以确保模型整个生命周期的安全。

当前正着力研发工具集成阶段的间接注入攻击防御新方法，包括调用链分析与输入数据过滤技术。基于标准化攻击测试集的智能体攻击鲁棒性基准测试体系加速发展，通过定期可靠性评估持续完善。同步构建面向非可信数据源的安全管控策略，具体措施包括建立隔离沙箱环境、实施域名白名单机制，以及部署内容自动化审核流程以确保输入模型前的数据安全性。

⚡ 挑战 5.5

应对人工智能滥用

人工智能模型可能被用于恶意目的，包括黑客攻击、社会工程攻击、制作深度伪造视频和虚假信息。这威胁着个人、组织和政府的安全。应对这一挑战需要开发数据保护方法（例如使用水印）、伪造检测算法以及反网络攻击系统。

4. 前瞻性研究任务

★ 任务 5.1

针对不准确或恶意数据（包括不准确的数据、受限访问的数据、以及法律禁止传播且违反社会价值观的数据）所带来的风险，正在研发和完善多种处理方法

这些方法涵盖数据清洗、异常检测、访问控制与法

律合规性检查等，确保数据质量和合法性，从而降低人工智能系统因不良数据引发的风险。

对于智能体AI系统，一致性应延伸至工具使用链，包括在执行工具调用之前评估意图的安全性和副作用。此外，还必须考虑开发过程中优化程序的影响，以及此类操作中各个知识领域的鲁棒性。

★ 任务 5.2

制定通用方法以确保人工智能工作的可解释性并提高公众信任度——可解释人工智能 Explainable AI (XAI)

目前正涌现基于形式化系统、逻辑与知识库构建透明人工智能方法的研究趋势。与此同时，针对可解释性系统的测试基准、评估指标、验证协议及技术要求也在加速体系化建设。

对于用户或审计人员来说，操作说明也很重要：为什么调用该工具，考虑了哪些政策限制，如何评估风险。

★ 任务 5.3

创建方法和基础设施，以确保人工智能系统的安全开发 (MLSecOps)

这项任务包括制定和实施构建安全开发基础设施 (MLSecOps) 的原则，开发防范数据集投毒和模型攻击的方法，以及开发用于识别第三方库漏洞的工具。目前正在创建基准测试，以正式确定人工智能系统的信任级别。

★ 任务 5.4

深度伪造内容检测与数字水印防护技术的开发

开发深度伪造对抗技术已成为数字安全领域的核心要务。随着生成式AI的普及，即便是不具备专业知识的恶意行为者也能制作出高度逼真的伪造音视频内容，这亟需发展高效防护手段。数字水印技术正展现出特殊价值——不仅能实现伪造内容的事后检测，更能对原始合法内容进行预标记，为真实性认证建立基础框架。此类解决方案在应对虚假信息与保护个人数据的场景中尤为重要，其中快速识别伪造内容的能力直接关系到用户权利与自由的保障。

在构建可信人工智能的理论基础方面，目前正在开展一项独立研究，旨在建立一套形式化的理论，以界定人工智能方法的适用范围及其可靠性标准。通用理论的构建将有助于从经验性的解决方案探索过渡到安全算法的系统性设计，并为新的基础研究奠定基础。

5. 重要结论: 专家意见

安全性和信任正成为现代人工智能的核心原则，界定着其在关键领域的应用边界——从医疗和交通运输到公共管理。重点在于使模型目标与人类价值观相符，确保抵御攻击的能力，保护数据，并开发透明的决策机制。现代人工智能安全范式超越了技术上的正确性，涵盖了整个开发生命周期：从数据集生成和模型训练到最终运行。

安全工程实践 (MLSecOps), 模型可信度验证和测试以及可解释性协议的发展，为负责任的 AI 部署奠定了基础，使 AI 能够以可预测和公平的方式运行。

构建可信可解释的人工智能不仅关乎技术进步，更具备制度建构层面的战略意义。当前正积极推进国际标准体系、审计规范与认证机制的建设，旨在确保AI系统实现透明度、稳健性及问责性。通过开发本质可解释的模型架构与决策事后解析工具，将有效增强公众信任，降低技术滥用与歧视性决策的社会风险。

从长远来看，这一方向将成为数字主权架构的关键要素，有助于形成负责任地使用人工智能的全球文化，并促进安全、创新型经济的发展。

这一方向对于人工智能的大规模应用至关重要。

27% 的课题涉及实现人工智能目标与人类目标的“对齐”，塑造正确的价值导向，为技术进一步发展与应用奠定基础。

方向六

专用人工智能
(Narrow AI)



方向六

专用人工智能 (Narrow AI)

1. 方向简介

专用人工智能（以下简称Task-Specific AI, Narrow AI）是专门执行具体、明确定义功能的一类AI系统。与致力于模拟人类广泛认知能力的通用人工智能（AGI）不同，专用人工智能专注于解决特定任务，从而在限定领域内实现高精度与实用价值。

其核心优势在于其当前的实际应用能力：从工业计算机视觉应用到政务服务的语言模型均已落地。凭借持续的经济效益及对社会与科学的直接贡献，专用人工智能已成为推动AI技术进步的主要驱动力。机器学习技术现已广泛应用于从软件开发、多智能体系统规划到医疗诊断等诸多领域，这充分体现了专用AI解决方案在工业与科学领域的重要影响力。

人工智能的未来发展方向并非是单一的：预计将从Transformer架构简单的线性扩展时代过渡到架构多样化的时代。与此同时，Transformer仍然是许多先进系统的基础：

当代开源大语言模型（Qwen, LLaMa）正普遍采用Transformer架构，且越来越多以“混合专家”模式（MoE）实现；计算机视觉领域的基础模型（如Segment Anything Model 2 (SAM-2), DINOv1-v3）同样是基于Transformer的架构构建的：AlphaFold3则将Transformer与更复杂的专用组件相融合。

与此同时，多种替代路径也在积极研究中，从状态空间模型（SSM）包括Mamba + Transformer等混合架构（如Nemotron、Granite）到神经符号系统、世界模型和图神经网络。

在实际应用场景中，专业化解方案正得到日益广泛的部署：从基于简单启发式规则的系统，到具备多语言支持与检索增强生成（RAG）等功能的进阶工具。科研领域则重点探索可穿戴设备（wearables）在人机交互领域的潜力——包括基于生理数据的压力监测，以及增强现实技术的集成应用（如Ray-Ban Display glasses）。

专用人工智能的研究对社会可持续发展和公平正义至关重要。模型可靠性的不足，包括易产生“幻觉”，限制了其在关键和高风险领域（如医疗、交通和法律）的应用。提高可靠性和透明度将增强公众信任，并确保敏感领域技术的安全集成。在社会领域，专用人工智能解决方案已开始实现公民请求处理的自动化，从而减轻员工的工作量，防止职业

倦怠。同时，该领域的发展有助于人工智能的普及化和技术的扩展，例如支持资源有限的语言，从而避免社会不平等的加剧。

在此方向上，划分了涵盖其基础的三个子方向：

1. 计算机视觉 (Computer Vision, CV)

计算机视觉技术已达到工业级成熟度。大约自2023年起，视频生成作为新兴领域开始蓬勃发展：商业项目陆续涌现，视频生成时长从数小时缩短至数分钟，通过新型架构与优化方法的持续迭代，生成质量不断提升。MagicTime模型依托通过精细标注的时序视频数据集（2000+视频序列）进行训练，在模拟物理逻辑过程方面表现尤为突出。构建基础模型已成为计算机视觉领域的核心方向——这类经过海量数据训练的系统，能够对用户自有数据进行高精度的视觉任务处理。

在任意目标的分割与跟踪领域取得了新进展，如Segment Anything 2、SAM等模型，以及支持自然语言条件下的目标检测方法，例如基于YOLOE、Grounding DINO等模型。还包括通过基础变换模型（如VGGT）对相机姿态和视频深度图进行估算，利用视觉问答（VQA）与图像推理（Reasoning）任务进行研究，并通过开放型视觉语言模型（如Qwen-VL）、以及可API访问的GPT-4o、GPT5等模型实现相关应用。

2. 自然语言处理 (Natural Language Processing, NLP)

大型语言模型正被广泛部署，例如在政府服务中自动处理公民诉求，这显著减轻了人工负担并有效预防了职业倦怠。通过多语言联合训练，模型能够利用不同语言间的语义共性提升效能，但需注意语言种类的过度扩展并不总能带来质量的提升。

在实践中，一些公司，例如 Cohere，已经放弃了27种语言，转而使用最重要的6种语言，这可能是因为这样既降低了数据收集成本，又使最终模型指标更加稳定。

3. 其他专用人工智能技术

除计算机视觉与自然语言处理领域之外，**各类专业人工智能技术**正持续发展。

在软件开发中，模型正被用于优化编译器；“人工智能在科学领域的应用”这一概念正在兴起，

特别鸣谢D. 尤金、L. 巴埃萨·耶茨、A. 图图巴利娜在该方向所做的工作。

以解决制药、物理、化学、生物和材料科学等领域的问题。

2. 当前发展方向概览

计算机视觉 (CV). 当代计算机视觉领域呈现出两大特征：视觉信息分析方法的效率发展，以及生成能力的突破性进展。基于文本的跨模态生成技术正快速发展，涵盖文生图、文生视频(text-to-image, text-to-video) 等方向，3D内容的创建与编辑方法也日益成熟。

在专家的讨论中强调：“**人工智能领域的生成式研究与实践应用**，主要体现在文本到图像、文本到视频(text-to-image, text-to-video) 等文本到视觉的转换技术”，并指出扩散模型在此类任务中发挥着核心作用。

医疗影像领域广泛采用自监督学习技术(self-supervised learning)，这种模型能够有效利用未标注数据，例如在放射学影像分析中实现精准诊断。最新进展体现在合规预测与风险控制技术的突破，以及将AI诊断报告无缝连接到临床工作流程——医生在AI辅助下进行决策，同时始终保持着对诊疗过程的全局掌控。

在工业与医疗场景中已实现了计算机视觉系统的规模化部署：智能制造中的自动化分拣流水线稳定运行，临床诊断则广泛采用»第二意见»系统辅助医生进行影像分析。新兴的图像信息提取技术不断涌现，以高光谱重建技术为例，该技术能通过图像获取精确量化数据，其应用范畴已超越纯视觉感知层面。

自然语言处理 (NLP). 自然语言处理领域正迎来重大突破。2023至2025年间推出的多任务语言模型(GPT-5, Claude, Gemini)在架构上实现显著飞跃，通过融合多模态理解、增强逻辑推理与小样本学习 few-shot能力，展现出前所未有的通用性。大型语言模型(LLM)已渗透至多个行业领域，特别是在公共服务层面，政府机构开始采用LLM自动处理公民诉求，既有效缓解人工压力，也有助于预防职业倦怠。专业化的自然语言处理技术在人文社科领域同样发挥着重要作用，例如通过古籍手稿数字化识别保护文化遗产，或基于历史文本语料追踪语言演变规律。

检索增强生成 (Retrieval-Augmented Generation, RAG). 将检索机制融入文本生成流程：模型在生成答案前会主动查询外部知识源。这种方法已成为当代大型语言模型的基石，既能持续更新模型知识体系，又可依托事实依据进行内容生成，从而显著降低»幻觉»现象的产生频率。

这种方法在瞬息万变的世界中至关重要，因为“如果模型是基于往年的数据训练的，我们自然无法获得最新的数据”。一个应用实例是“为 RAGFlow 招生委员会开发人工智能代理”，该代理成功地利用 RAG 模块提供多种语言的最新答案。这凸显出检索增强生成技术在教育、司法、公共管理等依赖实时信息更新的领域中应用的巨大潜力。

思维链 (Chain-of-Thought, CoT). 作为文本生成过程中的推理技术，思维链技术的核心在于引导模型先构建显性的逻辑推理路径，再生成最终答案。该方法显著提升了大型语言模型解决复杂问题的能力。值得注意的是，思维链效应在超百亿参数的大型模型中尤为显著，这种突现能力表明：当被恰当引导构建思维序列时，大型模型会展现出超越预期的推理能力。然而该技术仍面临模型幻觉与内在认知局限等挑战。

未来的发展方向包括整合外部工具和知识，以及开发能够自我学习和自我诊断的系统，从而模拟人类“思考、记忆和规划”的完美闭环。将语义网络或知识图谱集成到人工智能系统中，可以通过提供结构化事实并过滤掉不一致或错误的输出，来提高推理的可靠性。神经符号人工智能和 RAG 系统正在积极探索这一领域。

结构化输出 (structured output). 结构化输出技术旨在根据用户指令生成严格遵循规范的数据结构。该技术对商业应用具有关键意义，因其使大型语言模型能够与现有技术栈实现无缝集成。然而确保模型输出始终符规范仍存挑战：研究表明即便是最先进的LLM在处理复杂结构时仍会出现偏差，这表明输出格式的可靠性仍需提升。为提升结构化响应的稳定性，开发者正积极采用特制提示模板与输出控制技术。此类技术的成熟将有力推动跨行业创新与科研探索。

将机器学习与结构化规则、科学知识和本体论相结合（即所谓的神经符号人工智能）被认为是一种很有前景的发展方向。这种方法能够约束模型，并确保模型行为更具可预测性和可靠性，尤其是在需要高精度和可解释性的领域。

多智能体系统. 多智能体系统与自主人工智能代理正被广泛应用于实际场景：以高校招生办公室的多语言专业助手为例，这类系统不仅能流畅使用多种语言，还能从知识库中精准提取关键信息（LLM + RAG）。工具调用机制(Tool Calling)的出现，使语言模型能够通过调用预设软件工具来解决各类实际任务。这种技术架构的灵活性为大规模应用开辟了广阔前景，成为极具吸引力的研究方向。

例如，LLM在深网研究中已有成功的应用案例，其中语言模型通过调用搜索引擎，生成关于特定主题的报告。工具调用机制的灵活性使得LLM能够应用得更广泛，这为未来的研究提供了一个极具吸引力的方向。

3. 推动和制约该方向发展的研究挑战

⚡ 挑战 6.1

数据稀缺和合成数据局限性

高度专业化的计算机视觉和自然语言处理模型在特定领域面临着高质量数据短缺的问题。例如，在医学领域，罕见病诊断所需的样本采集十分困难；而在网络安全领域，获取关于攻击或生物特征反欺骗的可靠数据也并非易事。

合成数据生成技术正变得至关重要，但其长期影响仍存在不确定性。数据短缺制约着模型的质量与可靠性，削弱了其实际部署的可行性。若能找到用合成数据替代真实数据的方法，将为AI在复杂领域的规模化应用开辟道路。正因如此，合成数据集的生成方法正在快速发展，包括建立专门的模拟器在内的各种联盟，以交换医疗和工业罕见数据。

与此同时，合成数据质量验证方法的研究也在同步推进，旨在最大限度降低训练结果失真的潜在风险。

⚡ 挑战 6.2

为解决计算机视觉任务构建高效空间表征

关键问题是缺乏用于计算机视觉任务的通用方法来表示各种传感器的空间数据。

尽管多模态模型和三维表示技术（NeRF、高斯飞溅）已经发展起来，但现有方法无法有效整合语义特征，仍然满足高性能任务需求。

当前仍亟需开发自适应技术，这些技术应能有效整合多源异构传感器数据，并支持复杂空间推理能力——这正是实现现实场景实际应用的关键所在。

⚡ 挑战 6.3

计算机视觉系统的软硬件协同优化

为了实现计算机视觉系统的高精度工作（如测量技术和医学成像领域），不仅需要算法的发展，还需要传感器、光学设备和硬件的不断完善。没有综合性的优化，就无法实现这类系统的广泛普及。硬件与算法的协同优化能够将昂贵的原型转变为广泛应用的解决方案，加速计算机视觉技术在医疗和工业中的推广。

这一领域对于人工智能走出实验室至关重要。跨学科项目正在涌现，汇聚了工程师和人工智能专家。新型传感器和专用芯片正被研发出来，以适应特定的计算机视觉任务。

⚡ 挑战 6.4

资源有限的语言，包括稀有语言

世界上大多数语言在数字语料库中的代表性不足，导致它们在大型语言模型中也缺乏体现。这造成了数字鸿沟，并降低了自然语言处理技术的通用性。支持小众（罕见）语言对于人工智能的全球普及至关重要。

这一突破将大幅扩展技术普惠范围，并通过开创性教育文化服务的涌现，有效提升社会公平性。

为了应对这一挑战，目前正在开发一些旨在收集和标注资源匮乏语言的文本的项目。此外，人们也在探索合成语料库扩充和多语言学习的方法，包括从使用更广泛的语言中进行知识迁移。

⚡ 挑战 6.5

采用替代架构

尽管神经网络Transformer占据主导地位，研究人员仍在探索面向特定计算机视觉与自然语言处理任务的非传统技术路径。例如，目前正在研究的神经形态脉冲网络和光学神经网络。

虽然这些解决方案尚处于基础研究阶段，其成熟度还不足以与主流方法竞争，但它们为人工智能架构的演进开辟了全新方向。替代架构的发展有望突破现有模型的技术瓶颈，在提升能效的同时开拓前所未有的应用疆域。

⚡ 挑战 6.6

缩小机器人技术和自动驾驶领域中仿真与现实（Sim2Real）之间的差距

在机器人学领域，仿真与现实世界之间仍然存在差距（Sim2Real）。在仿真环境中表现优异的算法，面对真实世界时往往缺乏对物体真实物理属性的认知。要突破这一局限，不仅需要开发对环境变异性和传感器噪声具备鲁棒性的方法，更需构建能够感知和理解物理特性的系统。现有技术路线虽通过问答形式的数据集融入了物理先验知识，但在将这些知识直接迁移至实际操作对象时仍存在着显著的局限性。

机器人训练所需的高质量实景数据不足以及自治平台计算能力受限的问题加剧了这一挑战。在真实环境中确保机器人可靠且稳定运行仍是全球性难题。将可靠的物理特性信息整合进训练数据，是实现机器人可预测、安全和高效执行指令的重要步骤。

4. 前瞻性研究任务

★ 任务 6.1

开发计算机视觉技术以模拟现实世界场景（包括具身人工智能）

当前研究重点聚焦于开发能适应复杂现实场景的计算机视觉方法，旨在有效弥合仿真环境与真实世界之间的语义鸿沟(Sim2Real)。为实现这一目标，正积极推进领域随机化、仿真学习与主动学习等技术路径，同时发展视觉-语言-行动架构（Vision-Language-Action, VLA）——该架构使模型能够在模拟环境中通过自主行动积累经验，并将习得的知识有效迁移至物理世界。

这类技术体系对机器人技术、自动驾驶及工业自动化具有重要意义，因为这些领域不仅要求感知系统具备极高精度，更需确保行为决策的绝对安全性。

★ 任务 6.2

为各种计算机视觉任务（例如 SAM、Dyno v2）创建的基础 VLM（解决分类、检测、分割问题，类别列表开放）

这类模型需构建融合视觉、文本与传感器数据的统一空间表征体系。当前发展正沿着三个方向推进：基于对比学习与掩码重建的多模态预训练范式，二维与三维表征的协同建模，以及将物理合理性约束嵌入模型学习机制。

此类系统有望理解场景上下文，处理以前未知的物体，并进行空间推理，这将是向机器人和具身智能创建通用感知系统迈出的一步。

★ 任务 6.3

自然语言处理架构的高效训练与推理方法研究(涵盖自动化机器学习技术 AutoML)

研究重点在于通过自动化机器学习优化模型架构 AutoML，采用参数高效型微调技术，并系统整合知识蒸馏与推理加速方法——包括参数量化、网络剪枝及混合精度计算等核心技术。

目前，降低能耗和使模型适应分布式及专用计算环境正受到特别关注。解决这一问题将有助于创建更易于访问、速度更快、更新更便捷的语言系统，这些系统可应用于从智能助手到专业行业人工智能应用等各领域。

★ 任务 6.4

侧重于应用研究和开发推荐系统、语音识别和时间序列分析的架构的高效训练和执行方法(包括自动学习)

最后，一个重要的方向仍然是统一各个领域（推荐系统RecSys、语音识别S2T和时间序列分析TSA）中模型的训练和执行方法。

当前研究致力于开发能够高效处理流式数据、在保证预测质量的同时实现低延迟的模型架构。业界正广泛应用流式Transformer、状态空间模型与对比自学习方法，同时结合参数高效微调技术，使大型模型无需完整再训练即可适配垂直领域与特定编程语言场景。

结果将是创建通用架构和算法，使它们能够在资源有限的条件下实现自适应、节能和高精度，从而提升现代人工智能解决方案在各经济领域的可扩展性和普及性。

5. 重要结论:专家意见

科技界对专用人工智能的发展路径与优先方向展现出多元视角——从基础理论探索到产业化落地。

以下是专家提出的主要观点和分歧：

- 1) 统一化，依赖于基础语言模型，且额外学习量相对较少
- 2) 专业化，发展出特定领域的模型。

第二种方案通常能实现更高的系统可靠性，而第一种方案则具备更优的可扩展性与维护便利性。因此，在错误代价较高的领域宜采用第二路径，而对于容错空间较大的应用场景则更适合采用第一路径。

基于LLM的方法应用中的关键限制之一是其推理的不可信性(reasoning unfaithfulness)。尽管表面上看起来合理，文本推理往往不完整或扭曲地反映实际决策过程——忽略了影响决策的某些因素，将解释调整为看似合理但错误的答案，并误导性地呈现关键影响因素。因此，提升语言模型推理可信性的研究极具前景，这将逐步扩大其在高风险领域的应用范围，同时结合领域特定且更具透明度的方法。

众多专家强调人工智能应用研究的重要价值。特别关注“传感器-神经网络-处理器”架构的协同优化。这种科学驱动的机器学习方法论science-informed ML旨在缩小实验室原型与产业解决方案之间的实施鸿沟。实践层面意味着：需将基础研究与行业需求深度耦合，使新兴模型能够原生适配现实领域的的数据特性与硬件环境。

基础模型与高效空间推理的发展趋势。尽管现代基础模型（VLM）取得显著成果，其在推理能力、问题解答及三维空间物体定位等任务中仍存在明显局限。

开发能够进行高质量空间推理的模型和架构是现代的趋势，它可能为机器人、自动驾驶汽车、虚拟/增强现实系统、智能手机应用程序等创建高质量的空间理解系统(Spatial Reasoning)。

效能与精度。需要特别强调的是，必须开发更具效能的学习与推理方法。

计算机视觉任务被一些科学家视为一个独立的优化领域：需要在保持精度的前提下，实现计算机视觉模型的可接受速度和能效。这种平衡对于在资源受限设备中部署计算机视觉算法以及实时视频处理尤为重要。

低资源语言的问题。支持低资源语言的问题已被整个学术界公认为一个严峻的问题。在人工智能领域，人们认为有必要调整现有的语言学习模型架构，并生成新的语料库，以便尽可能多地涵盖各种语言。

构建更高质量的机器翻译模型仍是亟待突破的核心课题。这类系统不仅能用于生成多语言合成数据集，而且能通过提升自然语言处理技术的普及度与效能促进跨文明交流。若缺失此项突破，现有技术格局或将加剧语言资源分配的不平等，使多数语言社群无法享受现代技术红利。

超越神经网络的边界。一些研究人员呼吁应超越目前占主导地位的深度神经网络范式。有专家呼吁应采取更广泛的视角，积极研究人工智能的替代性模型（如符号模型、进化模型等），以避免忽略神经网络主流之外的宝贵方法。

对专用人工智能集成化发展的看法。一些专家认为，专业人工智能的进一步发展可能与基础模型的突破有关。谭舜泉提出：“我相信，自然语言处理模型若能取得真正的突破，任务6.3中所有问题都能迎刃而解。然而，要做到这一点，我们需要在基础自然语言处理模型方面取得突破。”

专家强调，基础语言模型质的飞跃将自然推动相关专用技术的协同发展。这种观点说明了该领域内部的一个分歧：一些人认为狭义人工智能的未来在于针对每个任务的专门模型，而另一些人则认为其未来在于基本模型的泛化及其随后对特定应用的适应。

应用最广泛的方向，包括对特定应用和行业至关重要的任务。

从科研格局演变视角观察，最具动态特征的领域当属其研究课题数量的持续增长与前景评估的流动性——这些变化直接受到外部环境波动的影响。

62% 的任务与计算机视觉技术相关

目前，深度神经网络范式占据主导地位。在该领域，更加广泛的视角同样必不可少，包括对符号模型、进化模型等替代性人工智能方法的关注与研究。

方向七

治理架构、决策机
制与智能体/多智
能体系统



治理架构、决策机制与智能体/多智能体系统

1. 方向简介

AI在管理决策与智能体系统领域经历了从规则驱动到自主认知的跨越式发展：早期基于固定逻辑的程序已演进为具备环境感知、自主推理与行动执行能力的复杂系统。这一变革主要得益于深度学习、强化学习（RL）及多智能体协作技术的突破。该领域的核心在于开发能自主决策、流程管控并与其他智能体交互以实现目标的人工智能系统，其优化范式往往着眼于长期收益最大化。当前应用疆域持续扩展，已覆盖协同机器人、集群智能、博弈对抗算法乃至跨行业复杂决策等前沿领域。

该方向的主要子方向如下：

1. 开发强化学习的算法

智能体通过与环境交互最大化累积奖励来学得最优策略，这是构建人工智能管理与决策系统的基础范式。

2. 智能体系统

作为具备感知、学习与自适应行动能力的完整自主AI实体，如今通过复杂架构设计已应用于最广泛的任务领域。

3. 多智能体系统

通过通信机制、竞争策略与集体决策等方法，实现多个智能体在复杂任务中的交互协作与协同优化。

历史上，用于管理和决策的人工智能始于经典控制理论与基于预设规则的专家系统。机器学习的兴起——特别是深度学习与强化学习的突破——标志着向数据驱动与自适应方法的重大转型，其中通过马尔可夫决策过程（Markov Decision Process, MDP）与马尔可夫博弈（Markov Games, MG）实现了“智能体-环境”及多智能体交互的建模范式革新。

近年来研究重心已转向大规模状态行动空间的可扩展性、现实应用中的安全保障机制以及泛化学习能力的系统性提升。当前显著趋势是构建具备自学习能力、可适应多元开放动态环境的智能代理，这类系统被期望能协调外部服务资源并自主进行任务优先级协调。

大型语言模型与基础模型在决策控制领域的集成，催生了旨在增强智能体推理、规划与交流能力的

新兴研究方向，为下一代自主智能系统的发展注入了全新动力。

AIRI人工智能的认知系统实验室首席研究员德米特里·尤丁提出：“近年来，机器人技术和自动驾驶领域出现了智能体的发展趋势，基础的VLA（Vision-Language-Action）模型如 $\pi 0$ 、Gemini Robotics、Gr00t等逐渐推动了通用机器人（能够作为家庭助手、自动化生产中的常规操作助手等）的研发。”

该方向发展的里程碑事件主要包括：

- DeepMind's AlphaGo 突破：在复杂策略环境中验证了强化学习的强大潜力，推动了通用博弈AI与管理系统的后续研究。
- 大语言模型与基础模型革命：LLM与基础模型的快速发展极大提升了智能体系统的自然语言推理、规划与交互能力，重构了多智能体技术范式。
- 机器人技术持续突破：机器人及无人系统在精细操作、敏捷移动与非结构化环境适应方面取得进展，推动AI从虚拟仿真走向实体应用，产生直接经济价值。
- 多智能体协作与联邦学习聚焦：分布式智能与隐私保护需求推动联邦多智能体强化学习在医疗、金融等领域的深入研究。
- 生成式AI智能体开发：LLM智能体框架与训练技术的成熟，加速了人工智能在各行业的规模化部署进程。
- 开发用于研究的智能代理，可以自动生成假设、进行实验、编写代码和撰写科学论文（例如 Google DeepMind 的 AlphaEvolve、Sakana AI 的 AI Scientist）。此类解决方案可以加速数学、计算机科学、工程等领域的科学发现。

决策控制与智能体人工智能作为核心发展方向，其突破将催生变革工业、经济与社会生活的下一代自主技术体系：

- 产业-交通-安全维度机器人学与自主系统的进步正加速无人驾驶车辆、无人机集群与先进机器人的部署，从根本上重塑交通运输、工业运营及高危环境作业模式；
- 生活品质维度智能体技术通过重构数字生态系统，推动自适应教育平台、临床辅助系统等个性

特别鸣谢Y. 布尔纳耶夫、A. 达尔维什、D. 尤金 在该方向所做的工作。

化服务发展，同时在网络安全领域增强威胁主动识别与处置能力，保障关键数字资产安全；

- 经济驱动维度通过提升生产效率、降低运营成本及创造新兴行业，AI已成为经济增长的核心引擎。制造业、物流管理与基础设施运维的高度自动化，预计将产生波及全球市场与竞争格局的规模性经济效应；
- 劳动力市场维度AI智能体的普及将推动就业结构转型——自动化技术逐步替代常规作业，同时催生AI研发、运维与监管领域的新兴岗位。这一变革虽可能造成部分职业被代替，但将通过大规模职业技能重塑计划创造新的就业机会。

2. 当前发展方向概览

当前人工智能在控制、决策、基于代理和多代理系统方面的发展现状：以深度强化学习（deep reinforcement learning, DRL）的快速进步和大规模语言模型（LLM）的日益融合为特征。深度强化学习已在复杂环境下的决策中展现出卓越能力，其应用范围涵盖游戏、机器人控制、资源管理和医疗保健等领域。

在过去一两年中，人工智能的发展方向发生了显著变化，其转向开发将自主学习、规划和交互集成到单一架构中的统一人工智能代理。

优先研究领域包括提高强化代理的鲁棒性、安全性和泛化能力。特别是，风险约束强化学习（risk-constrained RL）正在取得进展，其重点在于构建符合约束条件并最大限度减少故障的行为策略，这对于自动驾驶等高风险应用尤为重要。

此外，“代理人工智能”（将大型“基础模型”集成到代理行为框架中，从而发展出整体智能）的概念也日益受到重视。

多任务智能体学习方法正被用于创建适用于机器人、游戏AI和医疗保健等领域的自适应、多功能智能体。多智能体系统强调协调、通信和分散式决策，这得益于LLM的多智能体强化学习（MARL）和联邦学习技术的进步。基于LLM的智能体模型（包括Actor-Critic模型和基于角色的模型）的多智能体架构的开发是一个重要的发展趋势，并为需要合作与竞争的问题提供了可展的解决方案。

3. 推动和制约该方向发展的研究挑战

⚡ 挑战 7.1

多智能体系统的可扩展性与复杂性

多智能体强化学习（MARL）的关键挑战之一是处理庞大的状态空间和动作空间，其维度会随着智能体数量

和每个智能体复杂性的增加呈指数级增长。这个问题被称为“维度灾难”，它会显著降低现有方法在环境变得更加复杂时的有效性。然而，对多智能体架构的研究已超越工程范畴，旨在解决关于集体智能本质的根本性问题。多智能体系统中的相变和自组织现象或许能够为我们揭示，相对简单的智能体之间的交互如何涌现出全新的智能行为形式。

解决该问题将推动群体智能、协作机器人技术和大规模资源优化领域的突破。针对这一问题，研究人员积极探索hierarchical reinforcement learning（HRL），将复杂任务拆解为可管理的子任务。智能体越来越广泛地采用Retrieval-Augmented Generation（RAG）模块，大型语言模型（LLM）也通过代码和提示迭代改进模型。智能体设计变得更加主动，具备决策能力，包括对外部服务的调度和任务优先级的安排。开发了如MapGPT等用于去中心化多智能体规划的模型。自我对弈（self-play）成为显著趋势，AI模型通过相互交互生成大量数据，超过了基于人类数据训练的性能表现。

⚡ 挑战 7.2

人工智能智能体运行的安全性、可靠性与可信度

尽管强化学习（RL）在实践中取得了成功，但由于确保安全性、鲁棒性和可信度的复杂性，将这些成果转化为实际应用仍然充满挑战。在自动驾驶或工业控制系统等关键且高风险领域，这一点尤为突出。建立清晰的监管框架、明确责任界限、完善监督机制以及审计人工智能决策的方法，对于负责任地部署和公众接受更自主的人工智能系统至关重要。

社会对人工智能失误的反应存在着明显的不对称性：尽管自主系统在统计意义上可能比人类操作更安全，但一次故障就足以动摇公众对整个技术体系的信任。因此，攻克这一难题将为人工智能在交通运输、医疗健康及关键基础设施领域的实质性应用开辟广阔前景。

大量研究工作集中于风险约束强化学习及策略的形式化安全验证，利用分布式强化学习更全面地表征不确定性与风险，并通过带成本约束（cost constraints）的安全强化学习（Safe RL），借助上述标准的双准则优化来保障智能体的行为安全。

⚡ 挑战 7.3

具身智能体(Embodied Agents) 的模拟与现实之间的差距

虚拟仿真环境与真实场景间的知识迁移，始终是具身智能体发展面临的核心挑战（尤其是在机器人技术领域）。基于高精度仿真学习的行为策略，在物理世界部署时常因模型与现实的物理偏差、传感器

噪声干扰及环境不可预测性等因素导致效能衰减。这种“现实鸿沟”使得机器人在直接进行实体训练时既面临技术瓶颈又需承担高昂成本。

这个问题严重阻碍了智能机器人和物理自主系统的开发和部署。它需要在真实环境中进行大量风险较高的数据收集，并随后进行微调，从而限制了机器人研究和实际应用的规模化。克服这个问题将彻底改变机器人技术，加快开发周期，并拓展其在制造、研究和服务领域的应用范围。对于人工智能而言，这将促进跨领域自适应以及基于与真实世界有限交互的鲁棒学习方法的研究。

诸如领域随机化（domain randomization）的方法，通过在广泛范围内改变仿真参数，用于训练更鲁棒的策略，使其更好地泛化到现实中。研究元学习（meta-learning）和迁移学习（transfer learning），使智能体能够在最小的真实经验下快速适应现实世界的新环境。对物理动作的通用模型和视觉-语言-动作（Vision-Language-Action, VLA）多模态模型的研究，旨在构建更通用的表征，降低仿真模拟与现实之间差异的敏感度。

4. 前瞻性研究任务

★ 任务 7.1

构建融合文本与多模态信息的通用型智能系统：
视觉-语言-行动统一架构：Vision-Language-Action (VLA)

多模态系统，尤其是在机器人领域，将在未来十年成为研究的前沿方向。其目标是创建真正通用的多模态智能体模型，能够无缝地融合感知（例如视觉）、自然语言理解与生成以及物理动作（视觉-语言-动作，Vision-Language-Action, VLA模型）。此类模型必须展现出多任务处理能力、自学习能力，并且能够在开放、非结构化的环境中高效运行，而无需针对特定任务进行专门的预训练。

为实现该目标需开展跨模态大规模预训练，其中自监督学习、掩码自编码器与对比学习构成核心技术路径，这些方法被协同应用于视觉、语言及行为数据的整合流的对比学习。

强化学习将对于确保动态环境中的自我学习和适应至关重要，而层次化强化学习（hierarchical RL）有望使复杂行为场景结构化。此外，该任务存在数据不足的问题，潜在的解决方案包括生成合成数据和利用自我对弈（self-play）突破人为构建的数据集限制。

这一突破将催生适用于家政服务、工业制造与科研探索的高通用性机器人系统，同时推动能实现更自然的拟人化交互的智能个人助手的发展。

★ 任务 7.2

创建高效的方法，使智能体能够通过与环境交互来获取知识，从而实现目标

该任务致力于开发具有形式化安全保障的多智能体强化学习算法框架，确保系统在现实关键应用中满足稳健性与伦理合规要求。其核心在于超越纯粹的经验性效能，使多智能体系统能在预设风险边界内可靠运行，避免灾难性故障，并在复杂、不确定和充满对抗性的场景中始终保持行为可预测性。

★ 任务 7.3

智能体系统的研究与开发多智能体系统中相变现象的探索

目标是设计多模态架构，促进发达的集体智能的涌现，实现具有层次化组织与去中心化学习的一致行为，这些常常受到生物系统和复杂适应系统理论的启发。这包括开发“主系统”，能够动态创建和管理专门的窄领域智能体，以及研究随着智能体数量增加而出现的相变现象。

分层强化学习（Hierarchical Reinforcement Learning, HRL）将使智能体能够在不同抽象层次和时间尺度上进行学习。遗传算法与进化方法可用于在层次结构内实现智能体角色与通信协议的进化式筛选与优化。现代通信协议与多智能体系统中涌现的语言能够增强协同与信息交换能力。博弈论与机制设计则可被用于激励合作行为并管理竞争关系。

解决这个问题将有助于构建可扩展、灵活且稳健的多智能体系统，使其能够适应剧烈的环境变化，并解决远超单个智能体能力的问题。这将催生出用于复杂监视和搜救行动的先进无人机群、具备实时适应能力的智能交通管理系统，以及用于大规模建设或环境监测的高度分布式机器人系统。

5. 重要结论:专家意见

用于控制、决策以及基于代理和多代理系统的AI发展，其显著特征是向集成化、自主化和自适应智能方向转变。深度强化学习（DRL）的快速发展以及基础模型（尤其是LLM）的集成是关键趋势。这促成了具有高级感知、推理能力以及在复杂环境中跨多种模态运行能力的代理的创建。

研究重心已从基于规则的专用系统，转向具备环境动态适应能力的通用学习架构。现实任务日趋复杂的特性正推动多智能体解决方案的发展，持续促进稳健协调、高效通信与分布式决策等领域的研究。总体演进轨迹指向构建更通用、更智能且高度自主

的人工智能实体，通过提升系统的学习适应性与环境交互能力，有效弥合理论突破与现实部署之间的实施鸿沟。

此外，随着智能体之间（包括跨各种互联网平台）的交互，解决多智能体系统中的安全问题，特别是

串谋和协同攻击方面的安全问题，变得越来越重要。

关键因素不仅包括技术进步，而且还包括责任问题和监管框架的解决。该领域的成功将取决于研究界能否克服模拟中令人印象深刻的演示与现实中可靠运行之间差距。

重点关注自主系统的实用、高效和可靠的解决方案。

该方向的重要发展趋势与向集成化、自主化和自适应智能的转变密切相关。

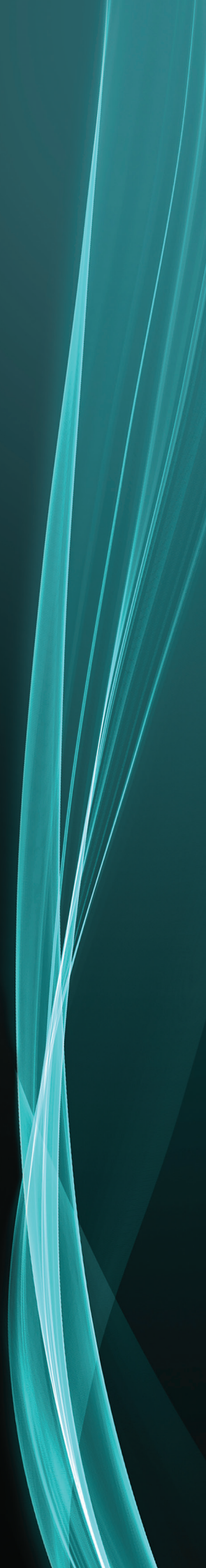
40% 的核心研究与强化学习
(reinforcement learning) 相关

智能体与多智能体系统正在迅猛发展，这与社会对自动化与自主化需求的持续增长紧密相关。对于众多任务类型而言，智能体可能演变为传统人工智能的新型进化形态。

在自动化与自主化需求持续增长的背景下，该方向面临的核心挑战是确保所开发技术与系统具备充分的稳健性与运行可靠性。

方向八

AGI要素





1. 方向简介

“通用人工智能要素”这一研究领域致力于推动人工智能从狭义任务向持续推理、终身学习、跨模态知识整合及复杂环境行动能力的系统演进。当前实践聚焦于构建未来系统的稳健性：可信推理与自我验证、环境交互中的持续学习、神经符号混合方法、具身智能与多智能体协同、脑启发架构等核心维度。

前瞻研究与专家访谈的共识表明：AGI的定义仍具流动性，其进展并非依靠单一突破，而是依赖多研究路径的融合汇聚。未来3-5年最具发展动力的领域预计集中在智能体/多智能体系统与持续学习方向；更长周期则看好神经符号计算与脑模型研究。

该领域的本质是从“工具型模型”向展现自适应通用性的系统转型——即具备可靠推理、规划、经验学习与跨领域知识迁移能力的体系。这些特性并非通过单一算法实现，而是通过推理反思机制、记忆与情境化系统、环境交互学习、混合知识表征及智能体行动等要素的协同作用形成。

该领域的界限在于：突破针对固定数据集的狭义优化，实现开放环境下的稳健运行——包括分布外泛化能力、决策可解释性与可验证性、长期记忆机制以及准确的自我效能评估。实用化评判准则体现为系统在资源与数据受限条件下，能否在保持安全可信的前提下解决科学探索、医疗诊断、工程设计等复杂现实任务。

目前主要的研究方向包括：

1. 推理与反思

强调提升推理质量（包括多模态推理）、因果理解、规划及自我校验机制——减少幻觉现象、置信度校准、推理链验证。这还包括可解释性实践和可认证的推理稳定性。

2. 持续学习

实现从静态训练到持续学习的转变：通过主动学习与课程规划、在线/离线强化学习、自我博弈及环境数据提取等技术手段，旨在突破固定数据集的性能瓶颈，实现在动态环境中的持续。

3. 混合人工智能

将深度学习与本体、知识图谱和逻辑相结合，以提高可解释性和鲁棒性。神经符号方法结合了工具使

用和知识检索（retrieval），提高了推理的准确性和可控性。

4. 具身智能与多智能体系统

通过真实/虚拟环境中的行动学习，实现多智能体的协同决策与群体通信。该系统以基础模型（LLM/VLM）作为感知与规划的核心支撑层，其关键挑战在于仿真到现实（Sim2Real）的迁移难题以及行为的安全性。

5. 大脑与心理建模

神经启发式模型是从脉冲神经网络到脑回路的模拟。它被视为提高能源效率、增强抗噪能力以及探索记忆/泛化新原理的长期发展方向。人工智能也被用于解读神经数据（例如脑电图数据），尽管这些数据稀缺且噪声较大。

从历史上看，人工智能领域经历了从符号人工智能（专家系统、本体论）到统计机器学习，再到深度学习和Transformer模型的演变——这一发展轨迹源于人们对通用模型的憧憬。然而，规模化应用暴露了其局限性（幻觉、反思能力弱、成本高），这重新激发了人们对混合性、持续学习和自主性的兴趣。在游戏中，自我对弈（例如围棋）展现了“从环境中学习”的潜力，而早期神经启发式思想则强调了生物学原理在提高效率和增强长期记忆方面的潜力。

过去五年间对该领域发展产生最显著影响的因素包括：

- **ChatGPT及后续LLM智能体**标志着基础模型潜能的规模化认知转折点，展现对话交互、工具调用与推理链能力，显著提升对AI可靠性、可解释性（XAI）及风险管理的标准，催生对通用人工智能要素的技术需求。
- **RAG/工具应用成为混合范式标准**：检索增强生成与外部工具/知识库的广泛集成，确立神经符号方法在产业界的实践地位，通过增强事实准确性与输出可控性，验证混合智能的技术价值。
- **思维链（Chain-of-Thought）与反思机制**：显性推理、自我验证与置信度校准技术的普及，重塑复杂推理任务的技术规范，推动可认证鲁棒性与幻觉抑制方法的研究进程。
- **具身（Embodied）/多智能体系统与训练环境**：支持智能体在动态环境中学习行动协调的平台

增长，实证从静态数据集向经验学习转型的可行性，使仿真迁移与行为安全成为核心研究议题。

- **脉冲神经网络和大脑神经网络是未来的发展方向。**人们对高效能且受生物学启发的神经网络架构，以及利用人工智能读取和解释神经数据重新燃起了兴趣。尽管存在噪声和据短缺的问题，但该领域极具发展前景。

2. 当前发展方向概览

在**推理与反思**子方向中，研究重心正转向提升输出可信度的技术路径：结构化推理链、自我评估机制、事实核查流程、可解释人工智能方法以及可认证鲁棒性实践。

在**持续学习**子方向中，在线学习与自我博弈机制正持续强化，业界积极探索通过经验获取与高质量合成数据生成来突破数据饱和瓶颈。

在**混合人工智能**子方向中，RAG/工具的使用以及与本体/知识图谱的集成实践已成为对准确性和可审计性重要任务的事实标准。

在**具身智能与多智能体系统**子方向中，行动学习与智能体协同训练平台正呈现增长态势，其发展主要受限于行为安全保障与现实场景迁移。

在**大脑与心理建模**子方向中，能源高效架构与神经数据解析方法的探索仍在持续深化，该领域尽管受数据可获取性的制约，仍保持稳健发展态势。

在过去一到两年中，对推理可靠性和减少幻觉的工作加强了；RAG和工具使用型智能体（tool-use agents）已在产品中得到应用；对多智能体系统和带反馈环境中的学习兴趣迅速增长。同时，效率议题逐渐形成：需要降低训练和推理成本，改善内存管理和个性化。

3. 推动和制约该方向发展的研究挑战

⚡ 挑战 8.1

推理可靠性增强与幻觉减少 (Reasoning and Reflection)

核心要点：可验证的推理链、置信度校准、对分布偏差的稳健性。范围：对医学、法律和科学至关重要；解决方案：可解释人工智能（XAI）、验证程序、反思循环和可认证方法。措施：制定评估和审计标准、基准以及行业实践规范。

⚡ 挑战 8.2

数据饱和与持续学习机制

当代人工智能系统正在从“训练-部署”的静态范式，向持续数据积累与经验演进的动态模式转型。

这种演进有助于突破固定数据集的性能瓶颈，并防止增量学习过程中的知识退化。其核心目标是使模型具备从环境中自主提取信息、丰富自身经验体系，同时保持已掌握技能的能力。这一方法论构成了构建具备终身学习能力的稳健长效智能体基础。为实现该目标，业界正采用在线强化学习（online RL）、自我博弈（self-play）、主动与课程学习等技术，辅以高质量合成数据生成机制，确保知识体系的持续更新。

⚡ 挑战 8.3

Sim2Real 仿真与现实的融合和具身/多智能体系统的安全

仿真环境至现实世界的技能迁移已成为核心挑战之一。对于自主系统与机器人技术而言，必须确保智能体在非受控环境中的行为安全性、可预测性与可复现性。当前研究聚焦于开发领域随机化方法、构建高保真仿真系统、建立测试场域及认证协议，通过系统化验证流程最大限度降低智能体在物理环境和人机交互中的行为风险。

⚡ 挑战 8.4

神经数据的缺陷与保护

神经启发式架构和脑电读取技术的发展受到神经数据稀缺性、噪声性和敏感性的限制。这些数据难以获取并难以进行匿名化，这阻碍了节能与认知驱动型模型的构建。应对这一挑战需要制定安全采集和匿名化标准，开发压缩技术，并构建紧凑的信号表示体系，从而在不违反伦理和法律规范的前提下使用神经数据。

⚡ 挑战 8.5

计算效能与可扩展性

模型规模的不断扩大和训练成本的不断攀升，使得能源效率和计算资源的优化利用变得至关重要。该方向的目标是在不牺牲质量的前提下降低训练和推理成本，确保技术的可扩展性，并保证通用人工智能（AGI）元素的可用性。

该方向包括计算过程的优化、新型硬件和架构的开发（包括分布式和节能方案），以及复杂计算系统中动态资源分配方法的构建。

4. 前瞻性研究任务

人工智能系统向更高自主性与复杂性演进的过程中，正催生安全保障、运行可靠度、可信构建与决策可解释性等基础性新课题。这些挑战亟需超越现有技术范式进行研究突破，其范围既涵盖机器学习的理论根基，也延伸至人机交互及环境适应的实践维度。

以下是根据专家讨论确定的一些有前景的研究任务。

★ 任务 8.1

通过持续学习提高人工智能模型的泛化能力和适应性

主要挑战在于克服现代人工智能模型的根本局限性：它们无法有效地推广到以前未见过的新领域或任务，除非进行范围更广的且昂贵的重新训练。

专家指出：“当我们针对特定领域训练神经网络时，并不能保证其在其他领域同样保持优越性能。”

需要开发“知识遗忘与保存的弹性”机制，使模型能够持续适应新任务，同时保留之前的经验，这被称为“无遗忘学习”

旨在解决上述任务，人们正在探索持续学习的方法、保留先前经验的新损失函数、领域适应以及基于人类记忆和遗忘原理的认知启发方法。

此类技术方案的实施将催生具备小样本学习能力与动态环境持续进化特性的通用型、稳健型自适应人工智能模型，这对机器人技术与个性化系统发展具有特殊重要意义。

★ 任务 8.2

推理可靠性：不确定性量化与幻觉抑制

当代大型语言模型虽能应对复杂提问，却常犯“极其低级的错误”，且易出现“幻觉”现象——即生成事实性错误但表述笃定的内容。

当前要务在于开发能使AI模型对其回答进行置信度评估与不确定性表达的技术机制，目标是训练模型学会“拒绝回答”或传递不确定性信号，这对保障其安全应用具有关键意义。

无监督方法利用模型内部统计数（熵、概率）来评估置信度，而监督方法则包括训练辅助模型来识别错误。此外，基于响应一致性分析（黑盒方法）、直接访问模型内部参数（白盒方法）以及训练模型来明确表达置信度的方法也在探索之中。这将有助于创建更稳健、更具自我批判性的AI系统，使其能够安全地与人类交互、拒绝可疑的响应，并提高关键领域推理的可靠性。

★ 任务 8.3

多智能体系统与复杂问题最优控制的发展

该任务旨在创建由多个相互作用的智能体组成的复杂人工智能系统。这些智能体能够在动态环境中协调、合作和适应。该任务超越了传统的强化学习（RL）。该任务包括开发用于采样、生成和微调大型模型的最优控制方法。需要建立可确保这类系统

可预测性和安全性的行为管理的理论和实践基础。

解决方法包括开发强化学习和动态环境最优控制算法。

这将推动构建能破解复杂难题、加速科研突破、提升运营效能，并实现大模型精准调控与优化的自主多智能体系统。

5. 重要结论:专家意见

该方向发展的本质特征体现为根本性范式转变——从构建专用型模型转向开发更具通用性、自适应能力与自主决策权的智能系统。

关键趋势是几个研究方向的融合：鲁棒推理、持续学习、混合方法和多智能体系统。

目前正在经历从基于固定数据集的静态训练，向通过与环境交互持续学习的范式转变，这一变革由现有数据饱和瓶颈与模型动态适应需求共同驱动。

使用混合神经符号方法（例如 RAG）以及外部工具来提高人工智能推理的准确性和可控性正变得越来越重要。

业界日益认识到，现有架构的规模扩展已显现其根本局限性——包括易产生»幻觉»与反思能力薄弱等缺陷，这使得不确定性量化成为当前最活跃的核心研究领域之一。作为回应，自我验证与置信度校准方向的研究正持续深化，这是构建可靠且安全的通用人工智能系统不可或缺的一步。

此外，多智能体系统和具身智能（embodiment）方向也在积极发展，基础模型被用作感知和规划的底层，主要挑战是实现技能从模拟到现实的迁移（Sim2Real）。

由于通用人工智能（AGI）的定义尚未完全形成，该方向的边界难以界定。

19% 的前沿研究涉及大脑与心理建模

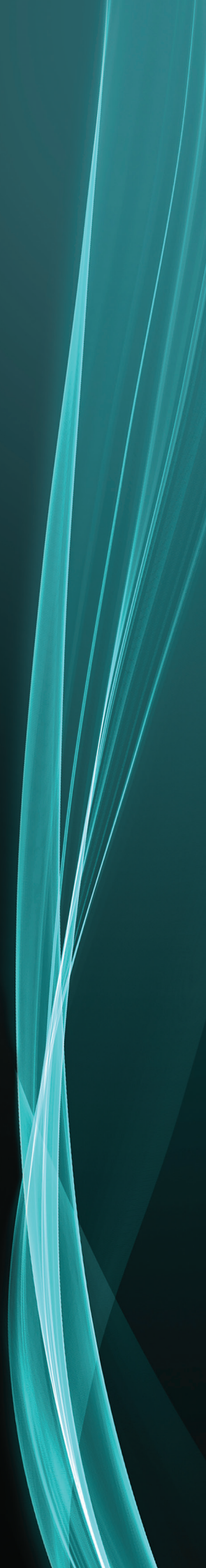
该方向发展的本质特征体现为根本性范式转移——从构建专用型模型转向开发更具通用性、自适应能力与自主决策权的智能系统。

该方向的可持续发展同样源于社会与监管部门对合乎伦理规范且安全可靠的人工智能系统日益增长的需求。

鉴于对自动化和自主化的需求，该行业面临的一大挑战是确保正在开发的技术和系统的稳健性和可靠性。

方向九

人机交互





1. 方向概况

人机交互（HMI）是一门跨学科领域，致力于研究、设计和实现人类与人工智能系统之间的通信接口及高效协作机制。其范畴已超越传统图形界面，涵盖所有形式的双向多模态信息交换，包括语音、手势、视线、触觉反馈，以及神经活动信号的直接读取解码与神经组织的刺激反馈。该领域的重要方向在于构建支持人类与AI深度融合及群体协同的工具体系。

目前，该方向分为以下三项子方向：

1. 与人类神经系统直接交互的技术载体

该子方向致力于研究与人体神经系统通过直接接触进行交互的系统。其研发内容包括脑信号读取系统的开发，以及为实现此目标所特制的、支持与神经组织进行双向交互的专用硬件与底层软件。当前研究重点关注开发生物相容性神经信号采集技术，旨在以最高精度和覆盖范围记录神经元群体的活动信号。这些技术与先进人工智能相结合，不仅将革新医疗康复领域，更将为自然智能与机器智能的全面融合奠定基础。

2. 传统人机交互的技术手段

本研究方向涵盖具有多模态交互功能的沉浸式环境（虚拟、混合与真实环境）开发，以及通过自然感官通道（视觉、听觉、触觉与嗅觉）实现反馈传递的技术设备研发。该方向还致力于探索人类行为处理与解析的技术算法，包括手势识别、情绪识别、语音识别，并延伸至用户意图预测及自适应接口的研发。

3. 人机交互的方法和算法

该子方向致力于构建方法论与算法体系，以保障人类与人工智能在前期两大方向技术设备框架内的协同工作。其核心在于开发增强人类能力的技术，实现人机团队的高效协作，涵盖协作机器人技术及决策过程支持系统。其中尤为关键的是对「脑密码」的破译研究，以及对侵入式/非侵入式双向脑机接口中大脑活动信号的解析。

施乐帕克研究中心（Xerox PARC）的研发成果带来了关键性的转变，苹果Macintosh电脑的普及进一步推动了这一变革，它引入了所见即所得（WYSIWYG）

的理念，并将鼠标确立为主要的导航工具。接下来的革命性转变是多点触控屏幕的广泛应用，其原理早在20世纪70年代就被提出，但由于iPhone的普及，才成为了现代触控用户界面（TUI）的标准。

语音识别技术的精进与作为现代助手算法核心的大语言模型发展，不仅实现了通过语音用户界面的自然对话，更显著提升了传统命令行交互的效率——尤其在编程和操作系统操作方面。

脑机接口领域的最新突破使得从脑活动中解码运动意图与语音信号成为可能，并实现了对神经过程的直接调控。这一进展既有望将人机信息交换提升至全新维度，也伴随着未经授权的语义内容读取和脑部直接干预所带来的伦理风险。

过去五年对人机交互领域发展影响最大的五个里程碑事件：

- **生成式人工智能**（尤其是大型语言模型，2022-2023 年）的爆炸式增长和普及，从根本上改变了人机交互的本质——从文本输入和点击转向自然对话。语音助手（Alexa、Siri）也得到了强大的升级，使语音交互更加有意义和有用。
- **侵入式语音脑机接口BCI（2024-2025）的突破性发展**，突显了该技术在实现与计算机直接交互方面的巨大潜力——包括对语音内容语义的精准解码能力。
- **Neuralink于2024年进入人体临床试验阶段**，即启动人类大脑芯片植入实验，这一里程碑事件极大地吸引了公众与投资界对侵入式脑机接口技术的关注。
- **元宇宙与扩展现实生态的发展、空间计算技术的规范化应用，以及非侵入式脑机接口领域的突破性进展**。Meta 和“元宇宙”（2021 年）的发布引发了沉浸式交互的技术竞赛，推动了 VR/AR 头戴式设备（Meta Quest、Apple Vision Pro）的发展，这些设备改进了眼部、手势和面部表情跟踪系统，以及触觉反馈和肌电活动解码技术（Ctrl-labs）。这些都是与嵌入用户物理世界中的数字对象进行交互所必需的。
- **人工智能伦理原则的整合（2019 - 2023）**：经合组织、欧盟等出台的倡议将透明性、公平性和责任机制（accountability）确立为AI信任人工智能监管倡议的启动（欧盟《人工智能法案》，2021 - 2024）：法律框架的建立直接塑造了

特别鸣谢K. 古格、I. 拉夫罗夫、A. 奥萨德奇、A. 萨夫琴科在该方向所做的工作。

人机交互的发展方向，要求其聚焦于稳健性、透明化与人类监督。

- 在COVID-19大流行（2020 - 2022）背景下这种要求尤显关键，该时期作为催化剂，加速了远程协作工具与人类存在虚拟化技术的普及应用。

人工智能监管举措的启动（欧盟人工智能法案2021-2024）：法律框架的建立直接影响着人机交互（HMI）的发展重点，需要重点关注稳健性、透明度和人工监督。在新冠疫情（2020-2022）的背景下，这一点尤为重要，因为疫情加速了远程办公工具的普及和人机交互的虚拟化。

随着人工智能与计算系统能力的增长，人机交互正成为关键瓶颈。若不能实现交互界面的根本性突破——提升其响应速度、可靠性与自然度——人类将无法充分释放人工智能的全部潜能。

从长远来看，HMI将开辟全新可能性：从增强人类认知的»计算扩展器«，到需要与大脑进行自然信息交换的医疗神经假体。然而该领域的迅猛发展也伴随着神经隐私保护、数据安全与技术滥用等重大风险。因此，发展可问责、可解释的人机交互系统不仅是技术议题，更是关乎国家安全与社会稳定的战略优先事项。

2. 当前发展方向概览

当代人机交互（HMI）发展正呈现人工智能、机器人技术与神经科技深度融合的特征，伴随而来的是交互范式的根本性重构。生成式AI与大型语言模型正在彻底革新人机自然语言交互模式。我们观察到从需要专门训练的交互界面，到基于世界模型的智能系统，再到能够预判用户意图的主动式多模态接口的演进路径。

神经技术领域的进展尤为显著：现代语音脑机接口（BCI）已能以接近自然的速度解码神经信号。仿生假肢在反馈界面的创建方面也取得了突破性进展。首批Neuralink植入物丰富了临床前实践，自适应深部脑刺激系统已在临床上用于恢复帕金森患者的运动功能。目前，人们正特别关注高时空分辨率脑活动记录技术的发展。能够以每平方毫米44个触点的密度记录皮层活动的设备已经问世，此类系统正在开展首次人体试验（Precision Neuroscience公司），用于解码此类信号的人工智能算法也正在开发中。利用纳米颗粒与大脑建立连接的技术为非侵入式、高精度地调节脑活动提供了可能，而这反过来又与其他发展趋势相结合，引发了一系列伦理挑战。必须克服这些挑战才能实现人机交互领域的可持续发展。

在这一过程中，可解释人工智能（XAI）是构建可信人机交互算法基础的关键所在。人工智能决策的可解释性正从一项附加功能转变为人机交互系统的

基本设计原则，从而确保决策过程的透明性，并增强人工智能与人类之间协作的推理能力，包括为逻辑推理过程提供可视化支持。

脑机接口（BCI）领域的突破：

- 语音脑机接口：实现接近自然语速的神经信号到语音的实时解码
- 仿生假体技术：成功开发具备触觉反馈的智能假肢，既支持与大脑皮层的直接交互，也能通过外周肌电信号控制并借助外周神经刺激实现精准操作首批接受 Neuralink 植入的患者以及公开演示光标/输入控制功能。
- 首例完全植入式无线脑机接口实现光标稳定控制与文本输入功能，这一突破性进展显著提升了学界对全植入式BCI技术的关注度，同时引发了对信号稳定性、系统安全性及康复应用场景等关键问题的深入探讨。
- 非处方助听器的独特类别。助听器市场的出现极大地降低了准入门槛，并刺激了用户音频界面（自助验配、与移动设备集成）的创新，为消费级神经和感觉设备树立了先例。
- 美国首款血管内脑机接口（BCI）由 Synchron 公司推出，属于新型微创接口，有效地连接了非侵入性脑电图和皮层阵列。

3. 推动和制约该方向发展的研究挑战

人机交互领域的快速发展带来了一系列挑战，这些挑战是当今研究工作应该着重解决的。

⚡ 挑战 9.1

实现可持续、稳健和同步的多模态融合

在多用户人机交互场景中，确保其融合的稳健性并能够考虑社交情境和用户意图至关重要。这需要有有效整合异步和异构数据（语音、视觉、手势）。多模态模型（视觉语言模型，VLM）的发展加剧了这一挑战的紧迫性。融合不同步会导致界面不可靠，且不可靠性会随着潜在冲突模态数量的增加呈指数级增长。这限制了人工智能在关键任务和敏感领域的应用。成功克服这一挑战将使我们能够创建“隐形”界面以及能够理解用户并主动采取行动的主动式、预测性界面。

⚡ 挑战 9.2

突破侵入式植入体的生物相容性与长期稳定性技术瓶颈

人体对植入物的免疫反应（胶质增生）会导致神经信号随时间推移而减弱，需要反复进行更换手术，

或根本无法长期使用。慢性侵入式脑机接口商业化和广泛临床应用的主要技术障碍使得该疗法具有潜在危险性和经济不可行性。目前，研究人员正在探索新型材料（柔性电子器件、水凝胶）、生物惰性涂层以及电极小型化技术，以降低免疫反应。突触神经接口也正在研发中，该接口通过在电极垫上形成天然突触来与神经组织建立连接。

⚡ 挑战 9.3

语义意图解码与运动指令解析的协同突破

大多数成功的脑机接口（BCI）能够解码运动指令（例如移动手的意图）。然而，如果不依赖运动关联，直接解码抽象思维、内心独白（意图）或情绪状态则困难得多。这限制了BCI在恢复高级认知功能（例如，帮助完全瘫痪的患者进行交流）和创建下一代交互界面方面的应用。目前，为了解决这个问题，人们正在利用低层次的神经网络（LLM）进行神经信号的上下文解释，并开展听觉和视觉意象解码方面的研（例如，语音BCI）。

⚡ 挑战 9.4

对大脑产生影响的编码原理研究

挑战在于破译大脑的基本“语言”——神经编码，它将信息转化为电化学活动模式。解决这一问题将使我们能够超越简单的脑活动监测，实现对神经元群的定向编程，从而为创建全新的系统铺平道路：从恢复丧失功能（视觉、听觉、运动）的医疗神经假体，到利用意念控制复杂系统的直接脑机接口。

⚡ 挑战 9.5

突破临床试验的行政监管壁垒

严格的监管要求与伦理审查机制，显著增加了侵入式脑机接口人体研究的实施难度。俄罗斯境内更因缺乏成熟的实践体系与配套基础设施，使得最需要参与临床试验的患者群体难以获得治疗机会。这种现状不仅延缓了科研进程与技术转化速度，而且使患者错失了获得突破性疗的潜在机会。国际层面正通过多措并举来破局：在顶尖高校附属医院（如麻总百瀚，Mass General Brigham）建立专科研究中心，制定针对»突破性疗法«的加速审批通道，并在临床试验设计环节吸引患者群体参加。

4. 前瞻性研究任务

★ 任务 9.1

开发能够理解用户需求与期望、感知其情绪状态的直觉式智能体

与传统接口不同，此类系统的交互基于目标设定原则——用户不指定具体动作，而是设定期望的

结果，代理自主选择实现这一目标的最优方式。为此，应用了多模态语音、手势、目光和生理反应感知模型，以及类似Vision-Language-Action和memory-augmented的代理等架构，能够考虑上下文、社会角色和用户的情绪状态。这类系统能显著降低认知负荷并缩短交互耗时，在机器人控制、自主系统管理与复杂数字环境操作等场景中具有特殊价值。前瞻性的设计应聚焦支持人机混合团队的动态协作网络——实现主动权的智能分配，开发能感知社会情境的交互模型，以及促进高效集体智慧的协同算法。

★ 任务 9.2

研究和开发双向脑机接口，破译“大脑密码”，并创建功能性脑图谱数据的基础模型

构建“意图—机器动作—反馈”闭环系统，结合多模态脑活动记录方法、基于foundation模型的自适应解码和强化学习算法，实现实时调整。关键要求包括确保延迟低于100毫秒、高解码精度，以及在非侵入式系统中提升信息交换速度。

★ 任务 9.3

建立人机界面评估的计量学基础

当前尚缺乏统一的评估标准来衡量用户信任度、认知负荷、参与体验及智能体自主性等关键参数。建立系统化的度量体系与测试规程，将为实现人机交互系统的效能与安全性提供客观评估基础拟使用的工具包括虚拟现实/增强现实沙箱、多标准指标（准确率、延迟、疲劳度）、生理指标和隐私测试方法。开发开放的人机交互基准和认证标准将为解决方案的可比性奠定基础，加速创新产品的市场准入，并增强人们对人机交互技术的信任。

5. 重要结论:专家意

人机交互正成为数字化转型的战略瓶颈：信息传递质量、交互速度、界面自然度及用户信任水平，共同决定着人工智能潜力的实现程度。这既适用于支持意识层面交互的智能系统，也适用在直接连接大脑与神经组织的接口技术——既可以为患者重建丧失功能，也能健康用户增强机能。

传统人机交互正朝着多模态和神经认知系统方向发展：文本、语音、视觉和手势的融合已成为实际标准。“隐形”界面和智能环境（环境/物联网）正在兴起，可以提供情境化且不引人注意的交互方式。一个重要趋势是从“单用户单机”模式向多用户场景、人机协作以及人机交互模式转变。

神经接口领域已获得一些具有突破性的成果：侵入式语音解码速度接近自然交流；语义解码已取得成

功尝试；通过外周神经刺激实现带有感觉反馈的假肢操控（基于肌电）；脑机接口介导的脊髓刺激应用；基于状态的响应神经刺激系统（RNS）已商业化用于治疗癫痫、僵直和震颤；超高密度脑电（约4K通道）被证明具有高度的信息性。

然而，能够提供充足信息量的神经活动信息，目前仍主要依赖侵入式电极系统——理想状态下这类系统可同步记录上千个独立神经元的电活动。该方向的可持续发展亟需开发新型植入式电极阵列，在确保长期生物相容性的同时，实现对数以万计分布在大脑皮层不同区域神经元的活动进行记录。通过构建突触接口实现与神经组织的精准对接，正成为极具前景的研究方向。

使用神经接口不仅是为了读取信息，更重要的是为了进行刺激，这对于闭合触觉/感觉反馈回路至关重要，赋予使用者自主感，增强假肢的功能，并降低使用假肢时的认知负荷。为了实现这一目标，必须采用复杂的上下文敏感神经网络编码算

法以及时间和空间上分布的刺激模式，从而提供尽可能自然的反馈。

光遗传学和热遗传学技术也正应用于高精度、空间选择性的神经组织刺激。此外，目前纳米颗粒正被是为与神经组织建立双向接触的一种微创替代手段。决定神经接口技术开发竞争力和可行性的重要组织因素是建立国内和国际法律框架，为使用原型神经接口系统对真实患者进行功能替代实验提供合法化。

人机交互技术的发展路径在于构建完整统一、可信可靠且自主适应的生态系统。这些系统的核心特征将体现在人机协同能力上：从能预判意图的动态交互界面，到实现生物层级深度融合的神经认知技术。未来的进步不仅取决于技术突破，还取决于在监管框架、伦理准则与网络安全三重维度的交叉领域取得系统性突破。创建技术不仅是工具延伸、而且是增强人类潜能的合作伙伴，是实现数字化转型全部可能性的核心使命。

数字化转型的战略“瓶颈”在于界面（接口）的质量、速度与自然性。这正是决定人工智能潜力能否得到充分展现的关键。

该方向最重要挑战是开发能够理解社会背景、角色和等级制度的模型。

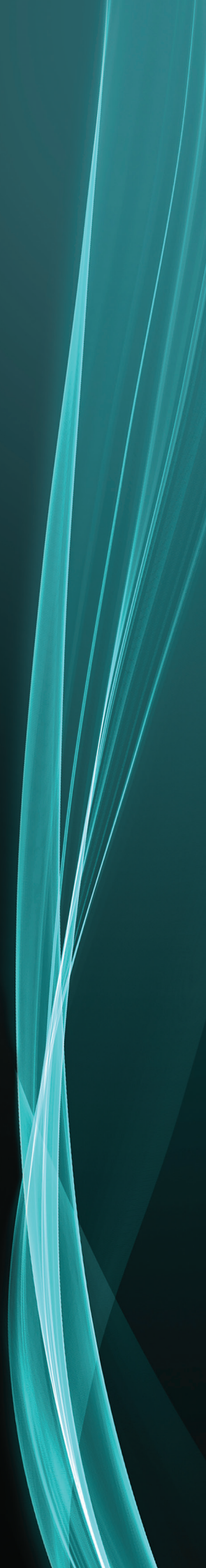
生成式人工智能与大语言模型领域的突破性进展，已显著改变了人机交互的基本范式与研究疆域。

对于自动化和自主化的需求，该方向面临的一大挑战是确保正在开发的技术和系统的稳健性和可靠性。

80% 的任务都与通过各种方式与人工智能协作来扩展人类能力有关

方向十

人工智能时代的社会





人工智能时代的社会

1. 方向概况

人工智能技术在带来显著优势的同时，也正对社会体系产生全面变革性影响。这种影响贯穿人类活动的各个领域：在经济层面体现为传统产业的结构变革与新兴“未来产业”的崛起；在社会层面引发生活方式的深度重构、就业模式的转型及教育范式的革新；在文化领域内容创作与传播机制发生根本性转变，推动文化生产逻辑的重新定义；在科学技术领域则催生了新的科研范式，标志着科学认知组织方式进入了新纪元。

正如马拉喀什卡迪·阿亚德大学高等技术学院计算机科学与数学系副教授阿齐丁·格扎兹教授所指出的：“未来数年人工智能的影响将引发新一轮的社会进化，这凸显出我们亟需通过部署AI算法提升生活质量，并与人工智能实现协同发展。”

专家在集体研讨中多次强调，技术突破往往伴随着重大社会与伦理风险，AI的应用速度往往超越了国际及国家层面适应性监管框架的建立进程。研究人员指出：

“我们正在等待‘全球社会契约’，在技术适用性达成共识的前提下推进技术建设。”

该方向将人工智能与社会的关系概念化为三元互动结构，涵盖治理体系、伦理规范与社会经济转型三大维度，强调这些领域具有相互增强效应，必须协同发展以实现以人为本的发展。

基于国际前瞻大会成果，该方向已形成了以下核心子方向：

10.1 全球人工智能治理机制，包括人工智能监管：

- **国家和全球人工智能治理体系。**近年来，大多数司法管辖区都致力于在国家和全球层面构建人工智能治理体系。人们反复强调并实践的结论是，人工智能治理必须超越国界，基于多边合作的全球框架。尽管各国在人工智能监管方面存在明显差异，但人工智能治理体系的进一步发展要求构建一个更加多元化和全面的范式，超越狭隘的国家监管模式。与此同时，许多研究人员指出，法律框架系统性滞后于技术发展速度，造成监管真空，进而加剧公众的忧虑和社会的紧张。然而，值得注意的是，在国家层面创建

人工智能监管沙盒的国际合作正在发展——从巴西到莫桑比克再到印度尼西亚，以及整个欧洲大陆——这有助于各国政府和创业生态系统在人工智能监管和创新之间找到最佳平衡点（欧盟计划即将发布相关的实施指南）。

- **国际合作。**过去一年间，人工智能领域的国际合作呈现显著的发展趋势。研究者强调，当务之急是开展广泛国际对话以及在AI技术治理与发展机制上达成共识，以推动安全、负责任且可信赖的人工智能技术的发展。该领域汇聚了全球各层级力量：从联合国大会主导的“人工智能治理全球对话”平台及科学专家组，计划每年召集各国共同制定安全可靠的AI系统开发规则到国家层面的重要倡议，其中值得关注的是中国提出的构建“世界人工智能合作组织”倡议，该倡议旨在确保各国平等参与并共同制定人工智能技术发展的国际规则。

10.2 人工智能伦理

在人工智能时代，伦理问题是社会讨论的核心。若不遵循诸如非歧视、算法的透明性和可解释性、数据保护、安全性和可靠性、问责制和监督等原则，可持续的技术发展将无从谈起。专家群体已达成了“设计伦理”的共识，即在系统开发过程中将这些准则嵌入其中。鉴于人们日益关注人工智能技术可能复制和放大的偏见和不公正风险，这一点尤为重要。人工智能伦理不应局限于程序性的公平和透明性概念，而应涵盖伦理的社会技术层面，包括算法公平性、认知多样性以及数据收集过程和系统设计中存在的权力结构等问题。

10.3 人工智能技术对社会影响的研究

- **经济与劳动力市场。**人工智能技术作为强劲的经济驱动力，在提升生产效率与重塑就业市场的同时，也潜藏着加剧社会不平等、职业替代效应及资本向大型企业集中的风险。科学界指出：“AI加速了生产进程，但未必能实现利益的公平分配”。国际讨论聚焦于利润的公正分配，提出自动化征税、收益再分配等补偿机制，甚至将全民基本收入视为缓解社会影响的潜在工具。对AI引发的经济转型分析，应透过技术政治经济学视角，考虑自动化如何重新分配价值、数据权利与议价能力，同时关注AI红利分配、数据劳动价值认定及价值创造公平分配机制等新兴议题。

- **文化认同。**在人工智能技术发展的文化层面，应特别关注人工智能技术的文化认同原则：语言和交流风格，基于历史和区域数据训练人工智能，反对文化标准化，礼貌和“禁忌”模型，政治和社会背景。专家强调，需保持文化多样性，提升数字素养和对人工智能产出的批判性认识。否则社会可能面临文化代码同质化和“后真相”现象加剧的风险。该方面可在数字主权和多元化背景下重新理解，从而强调语言和认知多样性在人工智能开发中的重要性，以防止算法统一化并保护数字时代的文化遗产。
- **研究人类在不同领域对人工智能自主性的接受程度。**研究人类作为人工智能系统用户在不同社会情境下对人工智能自主性的接受程度至关重要，因为这一参数决定了社会在医疗保健、交通运输和金融服务等领域将决策权委托给系统的意愿。这项研究应着重分析算法的技术能力、信任和透明度等心理因素以及风险水平和文化特征等情境变量之间的相互作用。

印度尼西亚泰尔科姆大学校长苏扬托教授提出：“在后真相时代，研究应致力于维护数据主权、人类尊严与文化自主性。”

社会与技术互动的历史表明，每一次重大的自动化实施，不仅伴随着生产力的提高，也伴随着严峻的社会挑战。19世纪的工业化催生了卢德运动，20世纪下半叶的机器人技术引发了关于就业的辩论，而21世纪初的数字革命则提出了关于隐私和数据控制的问题。这些例子表明，技术进步必然需要社会做出相应的调整——需要能缓解矛盾的社会、法律和文化机制。

人工智能技术发展的当前阶段也遵循着这一模式。最初的解决方案是在局部范围内选择性地实施的，逐步改变了整个行业。机器学习系统和神经网络的广泛应用不仅展现了其加速流程和提高效率的潜力，也带来了新的风险：

从算法歧视到隐私与透明度的威胁，从职业被替代到劳动力市场结构变革，从文化符号标准化到新型身份认同构建——为应对这些挑战，伦理准则、“设计即伦理”规范、人工智能影响评估标准及监管框架已开始出现，但这些措施均属事后形成的，证明了技术持续发展领先于制度建设。

在科学讨论中，专家们提出观点，认为未来应通过增强人的能力实现而非通过替代人来实现。因此，自动化和人工智能技术的应用历程遵循着循环演进的模式：技术飞跃开启了新的视野，但同时也提出了关于社会公平、监管、文化适应以及国际合作必要性的问题。

在影响该方向发展的关键社会经济事件中，以下事件尤为突出：

- 近年来，**生成式人工智能模型**在公共服务和消费产品中的广泛商业化和集成，正在形成一种全新的技术范式。尽管这种发展具有优化用户体验和扩展功能的潜力，但也带来了一系列远超纯粹技术问题的严峻挑战。这些挑战包括与内容创作相关的伦理困境、人工智能“幻觉”（即模型生成看似合理但实际上不可靠的信息）的根本问题，以及决策偏差的风险。
- **人工智能引发的风险格局**要求建立国家与国际层面的协同监管体系：
- **全球伦理规范制度化。**近年来，出现了从企业自发原则向行业标准、国家战略与国际公约的范式转变。从2020年代初期的“负责任AI”企业自律准则，至2021-2023年间已发展为政府机构间通过的具有约束力的政策框架。这些文件包括联合国教科文组织的《人工智能伦理建议书》（2021年）、俄罗斯的《人工智能伦理守则》（2021年）、经合组织和二十国集团原则等。过去五年，将伦理原则纳入国家战略和制定全球框架的趋势日益明显，这反映出相关讨论日趋成熟，并试图将价值准则转化为政治和法律机制。
- **立法进程显著加速。**2020至2025年间，全球针对人工智能应用的立法倡议显著增长。根据《斯坦福AI指数2025》报告，法律文件中提及AI的频率较2023年增长21.3%。国际社会正逐步构建人工智能技术法律治理范式，目前已形成三种典型监管模式：限制性监管、混合型监管与创新促进型监管。
- **发达国家和发展中国家之间的数字鸿沟日益扩大。**人工智能技术的普及程度不足，削弱了国家层面和国际层面的竞争力。大型企业在人工智能市场的主导地位对全球人工智能的整体发展产生了负面影响。发达经济体不断在工业、公共管理和教育领域应用人工智能，提高生产力和效率，而发展中国家则面临进一步落后的风险。由此产生的不平衡加剧了全球收入水平、教育质量和创新能力的的不平等，造成了一种数字殖民化现象，发展中国家不仅没有参与人工智能技术的发展，而仅是数据的提供者。
- **经济与就业结构变革。**人工智能对劳动力市场的冲击已成为近年最受瞩目的焦点议题。2022-2023年生成式技术的大规模普及加速了脑力劳动的自动化进程，引发关于职业未来的全球性辩论。前瞻研究与专家多次指出，数亿工作岗位将面临结构性重组或消亡，同时全新的就业形态也将应运而生。该方向的重要性在于它决定了社会对人工智能的信任边界，以及在创新与风险之间的平衡。国家和社会机构能否保障透明度、问责制和公平分配收益，将影响社会稳定、机构合法性和劳动力市场的可持续性。忽视这些问题可能导致歧视加剧、隐私被

侵犯和社会不平等加剧，而“伦理内嵌”与审慎监管的系统实施则有助于将损失最小化，为长期信任打下基础。在经济领域，完善的法律框架和社会责任标准有助于降低交易成本，使实施过程可预测，并减轻市场权利过度集中的风险。在文化和教育领域，在人工智能普及过程中保持身份认同和多样性，能够提高公众对这项技术的接受度，强化价值体系，并支持创新发展。总而言之，这种方法确保向人工智能时代的过渡不仅遵循效率逻辑，也遵循可持续性、公平性和人类尊严的逻辑。

2. 当前发展方向概览

许多专家对人工智能技术的监管滞后于其应用速度，无法有效应对其对社会造成的负面影响表示担忧。

近年来，人工智能领域的监管举措层出不穷，但这些举措都基于不同的基本方法。例如，欧盟《人工智能法》对人工智能技术的监管采取了较为保守的方式，引入了基于风险的方法，并对在欧盟境内部署人工智能模型的开发者提出了严格的法律要求。而包括中国和俄罗斯在内的其他国家则采取了更为灵活的方式，将监管与激励措施相结合，通过有针对性的监管限制及自我监管来实现。

包括英国、新加坡、韩国和日本在内的一些国家正在制定有利于创新的AI技术的监管模式，该模式最大限度地减少立法限制，并强调支持AI领域的科学研究和投资。

在一场学术研讨会上，深圳市人工智能产业协会执行秘书长比较了欧盟和中国的做法，并就人工智能技术的监管发表了意见：“有效监管的关键在于精准施策。欧盟侧重技术安全性、伦理合规性，建立了严格的规范体系，但客观上延缓了人工智能的发展进程；中国则在保障技术高速发展的同时，通过针对性立法保护关键领域技术应用，加速技术向实践转化，推动人工智能实现规模化发展。”

尽管各国在国家层面采取差异化治理路径，人工智能管理体系仍需突破国别监管边界，构建更具多维性与系统性的治理范式。

与此同时，从国际组织、各国立法机构到行业领军企业，都正在积极推进人工智能模型伦理评估工具的研发与应用，旨在系统性识别与修正算法偏见、完善数据保护机制、构建以人为本的AI模型，最终增强社会对人工智能技术的信任度。

一项国际前瞻研究的结果证实了这样一种论点：在实践中，人们采用了不同的方法来监测和评估人工智能系统，这导致了在人工智能系统行为的控制和责任问题上的分歧。

人工智能技术普及引发的经济转型不仅表现为生产效率的加速提升，更体现为资本向大型企业的明显

倾斜，这种趋势正在加剧社会的不平等。为应对这些挑战，全球范围内已掀起关于建立利益公平分配机制的深度讨论，包括征收自动化税与考虑全民基本收入等基本模式。

与此同时，在文化与社会领域，传统范式正在受到人工智能影响而被重新审视：创造力、教育过程甚至个人身份等概念都在发生转变。这些变化凸显了算法开发者的伦理责任，以及公民培养数字素养和批判性评价人工智能结果的迫切需求。解决上述复杂问题需要跨学科的方法，既要汇聚技术专家，也要整合人文领域的专业力量。

3. 推动和制约该方向发展的研究挑战

在发展该方向过程中，以下研究挑战尤为关键，既可能驱动创新也可能形成制约：

⚡ 挑战 10.1

构建人工智能技术治理体系，缩小人工智能领域的数字鸿沟

全球人工智能治理体系正处于初步构建阶段，这体现在各国政府、国际组织与企业层面涌现出多元化的倡议、立法提案与治理理念。当前核心任务是整合这些分散努力，通过国际合作建立基于平等、可持续性与可信赖的治理范式。专家特别强调，AI治理必须确保技术普惠性，着力缩小全球数字鸿沟。人工智能技术不应加剧现有的不平等，而应成为包括技术弱势区域在内的全球发展的催化剂。因此，AI技术发展需要与可持续发展目标及人类社会基本价值理念实现战略协同。

⚡ 挑战 10.2

国际合作

国际社会正逐步迈向建立公认的人工智能国际标准体系。此过程中的核心任务是组织广泛的跨国对话机制，以就人工智能治理达成全球共识。当前尤为关键的是实现从宣言承诺到实践行动的转型，切实保障各国在人工智能发展浪潮中的平等参与权。

⚡ 挑战 10.3

制定人工智能模型伦理特征的统一衡量标准

现有构成AI模型伦理评估方法学基础的度量体系，往往难以全面覆盖伦理合规的各个方面——从算法非歧视性与意外关联规避，到模型行为的全程可问责与可控性监督均存在评估盲区。

在这方面，专家强调建立综合评估框架的重要性，该框架允许对人工智能模型进行比较，从而建立一个普遍接受的评估标准，明确遵守人工智能伦理原则所需的一系列指标。

应该增加制定和制度化评估伦理后果及对人权影响的方法的必要性，以便将“默认伦理”原则转化为具体且可验证的实践。这些工具确保了在系统生命周期的各个阶段有效地贯彻价值观。

⚡ 挑战 10.4

失业和就业转型

日益扩大的涵盖智力劳动与创造性任务的自动化范畴，引发对就业市场未来的深度忧虑。诸多传统职业面临消亡的威胁，同时也出现了需要技能重塑的新兴就业形态。科学界面临的挑战在于精准预测其变革轨迹、构建适应性发展模型，并打造能应对新型劳动力需求的教育体系。

⚡ 挑战 10.5

文化多元性与身份认同

各国和地区的文化和伦理特征差异明显，这可能与制定通用国际标准相冲突。这种差异体现在语言、沟通方式、礼仪规范以及历史和政治背景等方面。挑战在于如何制定既能保护文化多样性，又能为人工智能的全球应用建立最低必要通用标准的方法。

将跨文化互动重要性提升至人类安全维度：大型语言模型承载的特定文化表征可能侵蚀文化认同。研究亟需明确如何通过数据过滤、标注与设计流程，确保人工智能技术既能维护社会联结与集体尊严，又能超越纯粹个体主义范式。

⚡ 挑战 10.6

复杂技术（包括人工智能代理）对人类认知功能的影响

新一代作为自主代理运行的人工智能增加了因失控带来的风险，加剧了系统决策过程中的不透明性。它们改变了沟通、授权委托的方式以及社会机构的结构。挑战在于研究这些AI代理的长远影响，制定框架以确保其安全和有益于社会的使用。

⚡ 挑战 10.7

探索人类对于人工智能在各个领域可接受的自主水平界限

人工智能技术的迅猛发展与通用人工智能的逐步出现，迫切要求我们明确界定自主系统决策权限的边界范围。

这项挑战旨在全面研究不同程度的人工智能自主性在人类各项重要活动中的社会可接受性。研究重点并非系统的技术参数，而是影响人工智能系统用户态度的复杂社会心理、伦理和文化因素。

4. 前瞻性研究任务

★ 任务 10.1

制定全球人工智能治理体系的方法

目前，人工智能领域的国际合作正稳步向前推进。

全球人工智能治理体系初步形成，其核心目标是将技术转化为解决全人类共同挑战的工具，而非引起新型冲突的源头。为实现这一愿景，需通过开发统一评估工具，系统整合各国法律体系及文化、社会与政治特征数据。此举将推动构建涵盖“全球南方”国家在内的包容性国际协定框架。

有必要建立跨文化伦理审计机制，以确保在脆弱地区部署的技术的公平性。其目标不仅在于确保技术普及，而且在于确保人工智能成果在当地环境中的有效性。

该任务解决方案的预期结果是在全球范围内达成人工智能的共识，或建立一个协调全球人工智能治理体系运作的国际独立机构。

★ 任务 10.2

建立国家人工智能技术监管体系

近年来，各国都在积极探索符合本国实际的人工智能监管模式。相关法律框架的建设高度重视文化特征、战略优先领域及现有法律体系的适配性。在此过程中，界定反映社会价值与伦理准则的“红线”——即法定的AI应用边界，已成为具有里程碑意义的关键环节。

预期结果应当是建立平衡的国家监管体系，既促进创新发展，又保护公共利益。该体系应协调融合国际标准与国家具体情况，营造可预测、透明的环境，支持人工智能技术的研发与应用。

★ 任务 10.3

人工智能社会伦理评估标准与度量体系的统一化建设，以及技术部署产生的社会经济影响测算

为解决这个问题，需要制定可比的指标和透明的框架，用于评估非歧视性、可解释性、隐私性、安全性和问责制。现有方法比较分散，使系统间的比较变得复杂，并降低了信任度。衡量对生产力、就业、收入分配以及市场力量集中度的影响。制定劳动力市场转型方案和公平分配福利的机制。

预期结果将构建统一的评估体系，该体系既能实现不同AI模型的性能对标，提升技术部署的透明度与可控性，又能建立认证制度基础，同时为国家政策提供经过验证的决策依据——包括监管与财税工具、教育战略及社会支持计划等全方位政策工具箱。

★ 任务 10.4

人工智能技术对社会影响的研究

新一代人工智能系统正日益增强其自主性，这加大了技术黑箱与行为失控的风险。

此类技术可能重构权力委托模式与社会机构运行机制。当前关键任务在于确定技术演进过程中需要保留的人性组成部分——这要求系统分析决策权让渡带来的心理与社会影响，并确立人机交互中的»拒绝权»原则。

本研究的预期结果，即研究人类对人工智能自主性可接受的限度，将为开发以人为本的界面和监管要求奠定科学基础，从而确保技术的创新潜力与保护公民的基本权利和安全之间的平衡。

5. 重要结论: 专家意见

鉴于人工智能技术的跨境性质，大多数专家学者一致认为，保持国际对话并就人工智能治理达成统一的全球协议，对于造福社会和促进人工智能技术的平衡发展而言，是最艰巨且最优先的任务。

值得注意的是，人工智能技术的发展伴随着透明度、问责制、数据保护和非歧视等基本伦理要求的不断完善。然而，这些要求的实施方法仍然零散且因国家而异，降低了实践的可比性，并阻碍了共同信任框架的构建。

社会经济影响日益凸显：自动化正推动劳动力市场结构的改革，加剧关于利益公平分配的全球讨论。当前核心议题涵盖自动化征税制度创新、新型社会保障机制构建以及职业技能重塑体系发展。

文化维度的重要性显著提升。全球技术既倾向标准化，又亟需尊重文化多样性、语言特性与社会规范。人工智能的社会接受度本质上取决于公民教育质量与公众对技术输出的批判性认知能力。

全球人工智能治理机制应建立在开放的全球对话之上，并将人工智能监管视为实现可持续共存的关键。由此产生的监管体系本质上应是社会技术性的，而非纯粹的技术性的，它必须认识到人工智能是人类选择的产物，并嵌入在现有的权力领域之中。

伦理学应当是一个集体反思的过程，而不仅仅是严格规则的集合。该方法应整合“道德多元主义”的原则，尊重文化和价值观的多样性，通过对话和协作形成共同的伦理框架。

构建一套全面的人工智能监管体系的挑战不仅在于选择一种（处理方法），而在于采用一种能够随着技术进步而演进的灵活监管框架。

总结

本研究被设想为“历史的一个片段”——一次大规模的尝试，旨在整合有关人工智能发展轨迹、其局限性和机遇窗口的知识。

我们始终将人工智能视为一个有机整体而非孤立技术趋势的集合：架构与算法、算力与能源、数据与法律、基础模型与生成式模型、安全与信任、专用解决方案与多智能体系统、通向量子智能的推理要素、人机交互界面，以及社会经济影响与监管框架——这十大要素共同构成了完整的技术生态图景。这种系统视角能清晰揭示制约发展的瓶颈、加速突破的动能，以及未来数年可扩展性边界的所在。

研究方法论依赖于广泛的国际合作：

我们举办了21场前瞻性研讨会，并与来自36个国家的100多位顶尖研究人员进行了访谈，同时结合开源分析和行业评论。这种学术视角与实践应用相结合的方式，使我们能够识别出共识领域（例如，能源消耗和数据短缺的紧迫性）和根本分歧（例如，代理的边界、可解释性的路径、架构优先级），更重要的是，能够识别出人工智能工程周期中的瓶颈。最终成果并非静态的“照片”，而是一张动态的地图，适用于定期更新和长期比较。

根据进行的大规模研究，在报告结论中我们提出了三个核心结论：

结论 1

如今，人工智能质量的发展轨迹并非由一个“神奇按钮”决定，而是在计算和能源成本不断上涨的背景下，由“算法-软件-硬件”的协同优化决定的。

结论 2

智能体与混合系统时代正在来临：我们将见证从被动辅助向主动AI范式的转变，以及向统计学习与显性知识、世界模型相融合的全新架构转变。

结论 3

进步的可持续性取决于“双重赤字”——能源和高质量数据；合成数据、主动学习和情境学习以及来自动态环境的新数据源不是奢侈品，而是必需品。

这项研究的实用价值在于它达成了科学家、工程师、企业和监管机构之间的共识。我们避免给出普适的解决方案，而是概述了一个决策框架：何时适合采用大规模基础方法，何时本地化和边缘人工智能至关重要；何时适合数据集中化，何时联邦学习更为合适；如何权衡质量提升与推理成本；以及如何衡量现实世界中人机交互界面的信任度和可解释性。

该框架有助于制定既保持现实可行性，又能在不同地域和行业间实现基准对齐的技术发展路线图。

我们将此次发布版本视为“实时”全球人工智能观察器的迭代版本。追踪各种变化——从优化器和压缩方法的改变，到多智能体系统的激增和新型加速器的出现——需要定期进行监测。

跨国协作对于报告撰写至关重要，因为能源环境约束、安全隐私标准、数据模型流通、工具协议兼容性等要素天然具有跨境属性。实验室、企业与监管机构的三方协同，能有效降低创新进程的交易成本，加速人工智能从理论科研向医疗健康、教育培训、工业制造与公共治理等领域的知识转化。

我们感谢所有参与者的专业且坦诚的讨论。本报告总结了已完成的工作，但并未对讨论做出最终结论。

我们诚挚邀请各国合作伙伴携手开启“新一阶段的协作”——共同构建知识共享空间、消除技术壁垒、验证创新假设，推动人工智能发展迈向负责任、高效益且真正普惠全球的新纪元。

附件

方向一
架构、机器学习算法、优化和数学

子方向	研究任务	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
1.1. 开发新的机器学习算法	1.1.1 机器学习与专家系统的集成（混合智能）											
	1.1.2 开发自监督学习方法 (self-supervised learning)											
	1.1.3 创建处理异构数据结构的通用方法											
	1.1.4 开发专用硬件系统和计算设备的算法											
	1.1.5 开发上下文和语言学习方法（语言强化学习、归纳学习、多智能体系统学习、反向传播）											
	1.1.6 开发多元低样本场景的算法（low-sample）											
	1.1.7 开发混合（经典与量子）算法及量子神经网络											
	1.1.8 创建机器学习模型特定领域质量评估的新算法											
	1.1.9 开发自进化人工智能算法											
1.2. 人工智能架构	1.2.1 创建深度网络架构的自适应方法											
	1.2.2 创建受神经科学和心理学启发的神经网络架构，包括脉冲神经网络											
	1.2.3 AutoML 方法的发展：元学习、自动特征工程、超参数优化、自动模型压缩等											
1.3. 计算加速	1.3.1 神经网络压缩方法的发展：量化、师生模型（teacher-student），网络剪枝等（network pruning）等											
	1.3.2 已知神经网络架构的计算优化（包括训练和推理阶段），神经网络的结构和物理特性（包括张量网络和矩阵网络）											
	1.3.3 开发加速计算的软件工具											
1.4. 分布式和联邦学习	1.4.1 大模型分布式联邦学习优化方法：通信开销削减与同步机制改进											
	1.4.2 开发分布式去中心化学习算法											
	1.4.3 开发联邦学习架构（包括混合架构）											
	1.4.4 联邦学习中数据处理与存储期间的非授权访问防护											
1.5. 人工智能的数学基础	1.5.1 研究降低机器学习模型复杂度的数学基础											



预期突破年是指预计在这一年中，将取得重大科学成就，从而从根本上推进某一特定问题的解决方案，并对技术发展和整个世界产生重大影响。



任务终结之年是指某个问题可以被认为已经解决的一年，并且预计该领域不会再有新的基础性或应用性发现。



方向二

人工智能计算



预期突破年是指预计在这一年中，将取得重大科学成就，从而从根本上推进某一特定问题的解决方案，并对技术发展和整个世界产生重大影响。



任务终结之年是指某个问题可以被认为已经解决的一年，并且预计该领域不会再有新的基础性或应用性发现。

子方向	研究任务	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
2.3. 机器学习和人工智能框架	2.3.1 计算优化（包括异构硬件上的计算优化）											
	2.3.2 创建用于训练和推理框架											
	2.3.3 创建符号人工智能和混合智能的框架（如SymbolicAI Framework, IBM Neuro-Symbolic AI Toolkit (NSTR), 经典符号人工智能工具（Prolog, 专家系统，语义网络/本体论），神经符号混合系统											
	2.3.4 构建基于代理的方案和应用框架（包括具身智能体）（如 AutoGen, CrewAI, LangChain, LlamaIndex, Semantic Kerne等											
	2.3.5 创建提示词工程框架（如 RTF, COSTAR, IEEI, APE, PECRA, OSCAR, TAG)											

方向三

人工智能数据

子方向	研究任务	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
3.1. 模型评估	3.1.1 创建维护基准测试，用于比较和评估神经网络的性能(CNN, Transformers, LLM 等)											
	3.1.2 创建人工智能评估指标											
	3.1.3 人工智能方法的在线评估（实际应用评估）											
	3.1.4 评估安全措施的有效性											
3.2. 数据生成、转换和维护	3.2.1 开发创建合成数据集的方法											
	3.2.2 主动学习与专家社群众包：从广泛人群中收集主观数据											
	3.2.3 研究数据组织方法（包括课程学习，curriculum learning)											
	3.2.4 用于数据生成、训练和在线模型验证的建模系统											
	3.2.5 开发数据质量评估、过滤、整理和排序方法											
	3.2.6 开发数据增强方法											
	3.2.7 创建用于训练大规模实用模型的开放数据集											
	3.2.8 开发纠正数据偏差和偏移的验证方法和算法											
3.3. 保障数据隐私与安全	3.3.1 开发差分隐私方法：向数据中引入噪声和失真											
	3.3.2 开发数据匿名化方法，确保敏感数据无法被复原											
	3.3.3 开发联邦学习中数据隐私保护方法											
	3.3.4 个人与版权数据的水印标记及训练中使用检测方法											
	3.3.5 生成内容的水印标记与检测方法											

- 预期突破年是指预计在这一年中，将取得重大科学成就，从而从根本上推进某一特定问题的解决方案，并对技术发展和整个世界产生重大影响。
- 任务终结之年是指某个问题可以被认为已经解决的一年，并且预计该领域不会再有新的基础性或应用性发现。

方向四

基础生成式模





预期突破年是指预计在这一年中，将取得重大科学成就，从而从根本上推进某一特定问题的解决方案，并对技术发展和整个世界产生重大影响。

任务终结之年是指某个问题可以被认为已经解决的一年，并且预计该领域不会再有新的基础性或应用性发现。

子方向	研究任务	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
4.5. 增强基础生成模型	4.5.1 研究开发短期记忆和长期记忆，以及随机活动无差别学习 (RAG) 和图随机活动无差别学习 (Graph RAG)											
	4.5.2 研究开发情境学习 (in-context learning) 和提示词工程方法											
	4.5.3 基础生成模型中工具使用 (如API Calls)											

方向五

安全、可信与可解释性

子方向	研究任务	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
5.1. 对齐 Alignment	5.1.1 开发降低错误和恶意数据相关风险的方法（包括不准确数据、受限数据（含法律禁止传播的信息）以及不符合社会价值观的数据）											
	5.1.2 开发识别算法足迹（footprints）的方法											
	5.1.3 创建在广泛应用的生成式（语言、多模态）模型中构建价值体系的方法（涵盖社会、文化和个体群体的价值观），包括协调多种偏好											
	5.1.4 开发对齐基准标准化开发方法的制定与对齐基准											
	5.1.5 推理模型的对齐											
	5.1.6 多模态对齐											
	5.1.7 机器学习的遗忘学习（unlearning）											
	5.1.8 开发人工智能模型的纠错方法											
5.2. 人工智能的可解释性	5.2.1 制定确保人工智能可解释性并提高其性能可信度的通用方法—— Explainable AI (XAI)											
	5.2.2 研开发事后解释（阐释）基于神经网络的不透明人工智能方法的方法											
	5.2.3 创建可解释和自解释模型 (Self-explainable AI)											
	5.2.4 基于概念和论证的人工智能 (Concept & Argumentation-based AI)											
	5.2.5 研究开发基于形式系统、逻辑和知识库的透明人工智能方法											
	5.2.6 制定人工智能系统可解释性测试的需求和协议											
	5.2.7 创建以人为本的可解释人工智能（Human-Centered XAI），包括考虑行业特性											
	5.2.8 确保大语言模型和多模态模型（LLM/VLM-XAI）的可解释性											
5.3. 保障人工智能安全开发与应用	5.3.1 开发确保人工智能系统安全开发的方法和基础设施（机器学习安全运维，MLSecOps）											
	5.3.2 开发防御数据集攻击（包括投毒和后门攻击）的方法和软件工具											

- 预期突破年是指预计在这一年中，将取得重大科学成就，从而从根本上推进某一特定问题的解决方案，并对技术发展和整个世界产生重大影响。
- 任务终结之年是指某个问题可以被认为已经解决的一年，并且预计该领域不会再有新的基础性或应用性发现。



方向六

专用人工智能 (Narrow AI)



 预期突破年是指预计在这一年中，将取得重大科学成就，从而从根本上推进某一特定问题的解决方案，并对技术发展和整个世界产生重大影响。

 任务终结之年是指某个问题可以被认为已经解决的一年，并且预计该领域不会再有新的基础性或应用性发现。

子方向	研究任务	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
6.3. 其他专用人工智能技术 （如S2T, RecSys, TSA等)	6.3.1 研究开发适用于 RecSys、ST 和 TSA 架构的高效训练和执行方法（包括自动学习)											

方向七

治力架构、决策机制与智能体/多智能体系统

子方向	研究任务	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035+
7.1. 开发强化学习算法	7.1.1 开发经典强化学习（RL）：Distributional RL, Risk-Constrained RL, Entropy-regularized RL, Safe RL等											
	7.1.2 制定包含多种环境和任务的预培训计划											
	7.1.3 开发复杂动态变化环境中强化学习策略Meta-Learning, Hierarchical RL, in-context RL等											
	7.1.4 开发离线到在线强化学习和实时学习											
	7.1.5 利用生成模型和智能体开发强化学习（智能体强化学习 Agentic RL)											
	7.1.6 开发真实环境和虚拟环境之间迁移学习的方法											
	7.1.7 开发自改进算法（自进化算法）(Self-Evolving Algorithms)											
	7.1.8 开发逆强化学习算法（逆强化学习）(Inverse RL)											
	7.1.9 开发高维场景和少量示例算法											
	7.1.10 无监督强化学习技术											
7.2. 智能体系统	7.2.1 创建物理智能体动作的通用模型：物理操作、多任务处理											
	7.2.2 创建整合文本处理和其他模态的通用多模态模型：视觉-语言-动作（VLA)											
	7.2.3 创建智能体能够通过与环境的交互来获取知识的有效方法											
	7.2.4 创建适用于虚拟和物理世界的基础智能体：具备多任务处理、自学习、在开放环境中工作以及使用开放工具列表的能力											
	7.2.5 开发从无标签行为数据中学习的方法（无动作学习）(Learning Without Actions)											
	7.2.6 开发自动化智能体设计方法											
7.3. 多智能体系统	7.3.1 开发基于核心模型的多智能体架构											
	7.3.2 智能体知识共享能力的发展：向其他智能体与人类学习											
	7.3.3 研究开发智能体系统：用于创建特定智能体的“主系统”、资源管理、鲁棒性、进化选择等											



方向八

AGI要素



预期突破年是指预计在这一年中，将取得重大科学成就，从而从根本上推进某一特定问题的解决方案，并对技术发展和整个世界产生重大影响。



任务终结之年是指某个问题可以被认为已经解决的一年，并且预计该领域不会再有新的基础性或应用性发现。



方向九
人机交互



预期突破年是指预计在这一年中，将取得重大科学成就，从而从根本上推进某一特定问题的解决方案，并对技术发展和整个世界产生重大影响。

任务终结之年是指某个问题可以被认为已经解决的一年，并且预计该领域不会再有新的基础性或应用性发现。



方向十

人工智能时代的社会



预期突破年是指预计在这一年中，将取得重大科学成就，从而从根本上推进某一特定问题的解决方案，并对技术发展和整个世界产生重大影响。



任务终结之年是指某个问题可以被认为已经解决的一年，并且预计该领域不会再有新的基础性或应用性发现。

最终报告的作者

编委会



亚伯拉罕·阿吉特

赛大学人工智能学院副院长兼院长，
印度



库兹涅佐夫·安德烈

FusionBrain实验室（AIRI）主任
俄罗斯联邦储蓄银行数据研究执行董事
俄罗斯



阿维蒂斯扬·哈鲁秋恩

俄罗斯科学院院士，俄罗斯科学院伊万尼科夫系统编程研究所所长
俄罗斯



梁正教授

清华大学教授，中政策研究中心副主任
中国



巴查宁·扎库拉·内博伊萨

辛吉杜努姆大学信息与计算学院，科学研究副院长
塞尔维亚



涅兹纳莫夫·安德烈

法学副博士，俄罗斯联邦储蓄银行以人为本人工智能中心总经理；
国际人工智能联盟秘书长
俄罗斯



巴埃萨·耶茨·里卡多

庞培法布拉大学
西班牙



奥塞莱德茨·伊万

物理与数学科学博士，俄罗斯科学院教授，AIRI人工智能研究所所长
俄罗斯



布尔纳耶夫·叶夫根尼

物理学/数学全博士，
斯科尔科沃科学技术研究院人工智能中心主任
俄罗斯



奥萨奇·阿列克谢

俄罗斯高等经济大学生物电界面研究中心主任
俄罗斯



布哈诺夫斯基·亚历山大

转化信息技术学院院长，人工智能研究中心“工业强人工智能”科学主任
俄罗斯



潘安辉

斯科尔科沃科学技术研究所正教授
越南



维齐尔特·尤里

俄罗斯国家航空系统研究所人工智能与机器视觉部主任、莫斯科物理技术学院人工智能研究所科学主任、物理与数学博士、俄罗斯科学院教授
俄罗斯



罗查·安德森

坎皮纳斯大学 人工智能实验室负责人



古格尔·克里斯托夫

g.tec医疗工程有限公司联合
奥地利



田野

安徽大学教授、博士生导师，计算机科学与技术学院
中国



加斯尼科夫·亚历山大

伊诺波利斯大学校长
俄罗斯



侯皓文

深圳大学广东省人工智能与数字经济重点实验室（深圳）副研究员
中国



达尔维什·阿什拉夫

开罗奥布尔科技大学（OUST）计算机科学与人工智能学院院长，埃及
开罗赫勒万大学理学院计算机科学教授
埃及



尤金·德米特里

技术科学副博士，AIRI人工智能研究所认知系统实验室高级研究员；
莫斯科物理技术学院人工智能研究所认知建模中心智能交通实验室主任
俄罗斯



阿哈托夫·阿克马尔

工程科学博士
沙罗夫·拉希多夫撒马尔罕国立大学主管国际合作副校长
乌兹别克斯坦



格沃兹德列娃·阿丽娜

俄罗斯联邦储蓄银行（Sberbank PJSC）以人为本人工智能中心专家
俄罗斯



阿尔梅达·克鲁兹·尤迪维安

教授
人工智能研究员、哈瓦那大学数学与计算机科学学院的教授
古巴



冈察洛娃·伊丽莎白

博士，多模态研究组组长，FusionBrain实验室，AIRI人工智能研究所
俄罗斯



阿塔瓦莱·维杰·阿南特

印度马哈拉施特拉邦瓦尔昌德理工学院校长兼 CSE 教授
印度



古格拉尼·阿卡什

印度数字基金会政策经理
印度



巴斯金·海姆

本-古里安大学电气与计算机工程学院副教授，INSIGHT实验室主任，数据科学研究中心研究员以
色列



格扎兹·阿齐丁

摩洛哥安全、智能系统应用和机器人技术（SISAR）研究团队负责人，卡迪·阿亚德大学
摩洛哥



贝洛-佩雷斯·拉斐尔

中央“玛塔·阿布鲁”·德拉维拉斯大学优秀教授，信息学研究中心主任，人工智能研究组组长
古巴



迪米特洛夫·丹尼斯

数据研究管理总监——Kandinsky基础模型管理部门负责人，俄罗斯联邦储蓄银行(Sberbank)
俄罗斯



博罗维克·鲁斯塔姆

俄罗斯联邦储蓄银行（Sberbank）人类精神人工智能中心人工智能转型负责人
俄罗斯



王殿辉

青岛科技大学数据科学学院院长、博士生导师
中国



伊达尔戈·丹尼斯·布埃多

高等教育部计算机化主任
古巴



叶尔绍夫·叶戈尔

工程学博士，俄罗斯人工智能研究院（AIRI）计算彩色摄影研究组组长
俄罗斯



布琼尼·谢苗

俄罗斯联邦储蓄银行（Sberbank）先进人工智能技术开发部总经理
俄罗斯



贾瓦德·穆罕纳德·阿里·穆罕默德

航空导航服务总公司。公司总部位于巴格达
伊拉克



布什科夫·尼古拉

生产力工程研发架构，研发中心，T-Technologies
俄罗斯



伊柳申·叶夫根尼

奥科公司人工智能总监，莫斯科国立罗蒙诺索夫大学计算数学与控制论学院信息安全系助教
俄罗斯



维里里·塞雷斯蒂娜

计算机科学博士。南非夸祖鲁-纳塔尔大学正教授，曾任数学、统计学和计算机科学学院计算机科学系主任
南非



伊里亚特·莱昂内尔

DATYS 高级研究员兼人工智能研究组组长
古巴



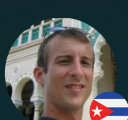
维图廖娃·叶利扎维塔

古马尔别克·道克耶夫阿拉木图能源与通信大学
阿尔-法拉比哈萨克国立大学
哈萨克斯坦



卡瓦列罗·莫塔·亚伊莱

卡马圭大学人工智能研究教授兼团队负责人；国际科学院院士
古巴



加西亚·博罗托·米尔顿

哈瓦那大学物理系复杂系统研究中心研究员
古巴



卡尔波夫·阿列克谢

俄罗斯科学院圣彼得堡联邦研究中心语音与多模态界面实验室主任
俄罗斯



科特查·凯坦

印度共生国际大学工程学院院长
印度



米洛瓦诺维奇·弗拉基米尔

克拉古耶茨大学工程学院电气工程与计算机科学系副教授
塞尔维亚



科努申·安东

物理数学科学博士，AIRI人工智能研究所Spatial AI研究组负责人
俄罗斯



蒙特西诺·佩雷纳·雷德尔

计算机科学大学（UCI）校长，电信与电子工程师，国家安全与国防专家，技术科学博士
古巴



科柳宾·谢尔盖

工程科学博士、教授、BE2R研究实验室主任、机器人与人工智能硕士项目主任、ITMO大学计算机科学与管理学院副院长
俄罗斯



莫拉莱斯-冈萨雷斯·安妮特

先进技术应用、模式识别与数据挖掘中心（CENATAV）高级研究员
古巴



科斯塔·若昂·皮塔

联合国教科文组织国际人工智能研究中心（IRCAI）创新经理兼高级研究员，以及巴塞罗那商学院（ENEB）人工智能相关主题讲师
UNESCO（IRCAI）



穆罕默德·阿卜杜尔·卡迪尔

阿里格尔穆斯林大学计算机工程系副教授
印度



克雷诺夫·亚历山大

Yandex俄罗斯人工智能技术开发总监
俄罗斯



穆罕默迪耶娃·迪尔诺兹

技术科学博士、教授；塔什干灌溉与农业机械化工程师国立研究大学
乌兹别克斯坦



拉皮娜·玛丽亚

俄罗斯北高加索联邦大学数学与计算机科学学院计算数学与控制论系副教授
俄罗斯



纳扎罗夫·法伊祖洛

撒马尔罕国立大学智能系统与计算机技术学院院长
乌兹别克斯坦



林欣

研究员
华东师范大学计算机科学与技术学院
中国



璘莫夫·阿列克谢

教授，俄罗斯国立高等经济大学人工智能与数字科学研究所所长
俄罗斯



马马托夫·纳尔齐洛

教授，技术科学博士、教授、塔什干灌溉与农业机械化工程学院（国立研究型大学）数字技术与人工智能系主任
塔什干信息技术大学
乌兹别克斯坦



尼科连科·谢尔盖

俄罗斯科学院捷兹洛夫数学研究所圣彼得堡分所高级研究员；圣彼得堡国立大学数学与计算机科学学院机器学习数学项目负责人
俄罗斯



曼苏罗娃·马迪娜

哈萨克斯坦阿尔法拉比国立大学信息技术学院人工智能与大数据系主任，物理与数学科学副博士，副教授
哈萨克斯坦



尼拉坎丹·苏布拉马尼

R. M. K. 工程学院智能神经认知知识与数据分析实验室（SNKDIR）教授
印度



马尔科夫·谢尔盖

俄罗斯联邦储蓄银行人工智能技术开发总监
俄罗斯



奥贾·阿帕拉吉塔

阿帕拉吉塔·奥贾
贾巴尔普尔PDPM IIITDM计算机科学与工程教授
印度



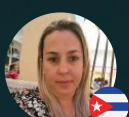
马西莫·梅切拉

正教授
意大利萨皮恩扎大学信息工程、计算机科学与统计学院，计算机、控制与管理科学与工程系；安东尼奥·鲁贝蒂大学（DIAG）
意大利



奥拉万德·达拉莫拉

比勒陀利亚大学信息学系研究教授
南非



门德斯-巴斯克斯·海迪

国家技术研究中心（CENATAV）生物识别部门负责人兼研究主任
古巴



帕诺夫·亚历山大

物理与数学博士，人工智能认知系统实验室主任，AIRI研究所
俄罗斯



帕特尔·萨姆普里特

助理教授（计算机科学与工程），瓦尔昌德理工学院：印度马哈拉施特拉邦索拉普尔
印度



鲁斯塔莫夫·萨米尔

计算机科学博士，副教授，ADA大学
阿塞拜疆



皮涅罗·佩雷斯·佩德罗·约巴尼斯

东方中央大学人工智能研究、发展与创新中心（CIDIIA）副主任
多明尼加共和国



萨夫琴科·安德烈

教授，技术科学博士，俄罗斯联邦储蓄银行（Sberbank）人工智能实验室学术主任
俄罗斯



皮亚德·莫尔菲斯·亚历杭德罗

亚历杭德罗·皮亚德·莫尔菲斯
计算机科学家，哈瓦那大学全职教授
古巴



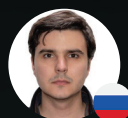
萨吉德·穆罕默德

穆罕默德·萨吉德
阿里格尔穆斯林大学计算机科学系助理教授
印度



波戈相·爱德华

亚美尼亚国家科学院信息与自动化问题研究所教授，技术科学副博士，认知算法与模型方向负责人
亚美尼亚



萨姆索诺夫·谢尔盖

国立高等经济大学国际随机算法与多维数据分析实验室主任
俄罗斯



普丽缇·卡纳 a

PDPM印度信息技术、设计与制造研究所计算机科学与工程教授，赞助研究与咨询项目院长
印度



萨菲特里·斯里

集团企业转型领导者、求知若渴的学习者、全球演讲者、新锐作家
印度尼西亚



皮埃拉·富恩特斯·艾伦

医学博士
哈瓦那信息科学大学研究员
古巴



塞佩罗·奈玛

哈瓦那技术大学（CUJAE）计算机工程系教授；人工智能研究团队负责人
古巴



陈鹏

安徽大学计算机学院/物质科学与信息技术研究院教授，博士生导师，中科院合肥智能机械研究所研究员
中国



斯克林尼克·阿列克谢

物理学和数学博士，人工智能研究所认知系统实验室强化学习智能体研究组组长
俄罗斯



拉纳·法亚兹·艾哈迈德博士

人工智能技术中心（AITEC）高级研究员（人工智能方向）
巴基斯坦



苏迪亚纳·多迪

印度尼西亚大学工程学院电气工程系教授（DoEE UI）
印度尼西亚



雷切尔·塞缪尔

德布雷伯汉大学可解释人工智能与自然语言处理主任
埃塞俄比亚



苏莱曼诺夫·易卜拉欣

技术科学博士，国际工程技术大学智慧工程技术系教授
哈萨克斯坦



拉希多夫·阿克巴尔

技术科学博士，撒马尔罕国立沙罗夫·拉希多夫大学人工智能与信息系统系副教授
乌兹别克斯坦



桑布瓦尼亚姆贝·姆布尤

南非大学计算机科学系主任
南非



罗德里格斯·菲格雷多·埃克托

副总裁 / 哈瓦那科技园股份有限公司
巴人工智能联盟网络代表
古巴



托拉尔巴斯·埃斯佩莱塔·拉斐尔·路易斯

哈瓦那科技园总裁，古巴人工智能联盟网络代表
古巴



罗戈夫·奥列格

物理数学科学博士，AIRI人工智能研究所可靠与安全智能系统研究组负责人；AIRI-莫斯科通信与信息技术大学SAIL实验室主任
俄罗斯



图图巴利娜·埃琳娜

计算机科学博士，俄罗斯科学院人工智能研究所（AIRI）应用自然语言处理研究组组长；俄罗斯科学院伊万尼科夫系统编程研究所高级研究员
俄罗斯



图达尔科夫·丹尼斯

俄罗斯科学院圣彼得堡联邦研究中心语音及多模态交互实验室主任
俄罗斯



哈密多夫·穆尼斯

信息技术副教授，人工智能研究实验室主任
撒马尔罕国立大学
乌兹别克斯坦



乌特金·列夫

工程科学博士，彼得大帝圣彼得堡理工大学人工智能技术高等学院教授
俄罗斯



埃尔南德斯·埃雷迪亚·亚尼奥

工程学博士
信息科学大学首席执行官
古巴



费布尔斯·埃斯特拉达·艾琳

技术科学博士，通信部副部长
古巴



侯赛因·埃萨姆·埃萨姆

他是埃及米尼亚大学计算机与信息学院的副教授，也是计算与人工智能研究小组（CAIRG）的创始人和主席
埃及



法济洛·沙夫卡特

理学博士、教授、数字技术与人工智能发展研究所生物识别系统实验室首席研究员
乌兹别克斯坦



陈昌盛

深圳大学副教授，
深圳北理莫斯科大学工程系教授，博士生导师，IEEE/CCF/CSIG高级会员
中国



费雷尔·明格斯·冈萨洛

副教授
移动机器人实验室主任
斯科尔沃科学技术学院
西班牙 / 俄罗斯



切卡利娜·维多利亚

技术科学博士，AIRI人工智能研究所FusionBrain实验室多模态研究组高级研究员
俄罗斯



秦飞巍

杭州电子科技大学计算机科学与技术学院副教授
中国



什皮尔曼·阿列克谢

俄罗斯联邦储蓄银行（Sberbank）人工智能科学中心总经理
俄罗斯



哈姆达莫夫·鲁斯塔姆

乌兹别克斯坦塔什干数字技术与人工智能发展研究所
乌兹别克斯坦



埃斯特韦斯·拉姆斯·埃内斯托

哈瓦那大学正教授，古巴科学院院士
国际晶体学联合会数学与理论晶体学委员会委员
古巴



拉夫罗夫·伊戈尔

医学博士，梅奥诊所罗切斯特分院脊髓神经调控研究组负责人
俄罗斯



阿伦·A·加罗法洛·埃尔南德斯

哲学博士，Awangio SRL 产品总监
古巴



阿斯拉尼扬·列冯

物理数学博士，教授，亚美尼亚国家科学院通讯院士，亚美尼亚国家科学院信息与自动化问题研究所离散数学系主任
亚美尼亚



赫玛姆·里萨

KORIKI（印尼人工智能协会）主席；印尼在人工智能联盟网络（AI Alliance Network）的代表
印度尼西亚

谨向参与本项目的戈尔班·亚历山大·尼古拉耶维奇致以诚挚谢忱。亚历山大·尼古拉耶维奇先生作为卓越学者，其开创性研究为现代科学发展作出了不可磨灭的贡献。

他的研究领域十分广泛，涵盖统计物理学、非平衡热力学、机器学习和数学生物学等

永垂青史，光照千秋



戈尔班·亚历山大·尼古拉耶维奇

中央大学及AIRI人工智能研究所人工智能、数据分析与建模实验室主任
俄罗斯

最终报告团队



涅兹纳莫夫·安德烈

俄罗斯联邦储蓄银行（Sberbank）以人为本人工智能中心管理主任，法学博士；国际人工智能联盟秘书长



查切·埃尔维拉

俄罗斯联邦储蓄银行（Sberbank）以人为本人工智能中心执行主任



阿尔秋金·奥列格

俄罗斯联邦储蓄银行（Sberbank）人工智能倡议实施与推广局执行主任



丘里洛娃·达莉娅

俄罗斯联邦储蓄银行（Sberbank）以人为本人工智能部门负责人



博罗维克·鲁斯塔姆

俄罗斯联邦储蓄银行（Sberbank）以人为本人工智能中心人工智能转型负责人



福基纳·索菲亚

俄罗斯联邦储蓄银行以人为本人工智能中心负责人



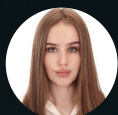
利索夫·阿尔谢尼

“人工智能发展支持与发展战略机构”项目办公室主任



泽姆佐娃·尤利娅

俄罗斯联邦储蓄银行（Sberbank）人工智能倡议实施与推广局首席数据研究员



格沃兹德列娃·阿丽娜

俄罗斯联邦储蓄银行（Sberbank）以人为本人工智能中心专家

致谢

编辑和组织团队谨此向以下人士致谢：

我们衷心感谢俄罗斯联邦政府和俄罗斯联邦经济发展部对本项目实施的支持，以及以下人士对俄罗斯人工智能科学发展做出的宝贵贡献：



切尔尼申科德米特里·德米特里
俄罗斯联邦政府副主席



科列斯尼科夫·马克西姆
俄罗斯联邦经济发展部第一副部长



索特尼科娃·奥克萨娜
俄罗斯联邦经济发展部战略发展与创新司司长

我们衷心感谢人工智能发展战略机构（SAPFIR）为本项目提供的支持，具体人员如下：



索尤兹诺娃·塔季扬娜
“人工智能研发支持与规划战略机构”项目办公室主任



泽姆利亚努欣·基里尔
“人工智能研发支持与规划战略机构”项目办公室执行主任

人工智能联盟协会对本项目的支持，特此感谢以下代表



瓦莱里娅·沃罗比约娃
人工智能联盟协会总经理

我们谨此向俄罗斯联邦储蓄银行（Sberbank）致谢，感谢其组织和支持本项目，具体负责人如下：



韦迪亚欣·亚历山大
俄罗斯联邦储蓄银行股份有限公司管理委员会第一副主席



别列夫采夫·安德烈
俄罗斯联邦储蓄银行（Sberbank）高级副总裁兼技术开发部门负责人



埃雷门克·马克西姆
俄罗斯联邦储蓄银行（Sberbank）副总裁兼人工智能和机器学习开发部主任



阿韦尔巴赫·弗拉基米尔
俄罗斯联邦储蓄银行人工智能计划实施与推广部高级总经理

致谢

编辑和组织团队谨此向以下人士致谢：

感谢AIRI人工智能研究所的科学家们参与最终报告的撰写工作，其代表如下：



布罗伊特曼·亚历山德拉
AIRI人工智能研究院市场营销与传播总监



佩夫娜·伊丽莎白
AIRI人工智能研究院公关经理

感谢俄罗斯联邦储蓄银行康定斯基生成网络团队创作插图：



尼库琳娜·塔季扬娜
俄罗斯联邦储蓄银行康定斯基基础模型部主管

报告设计



奇米尔·奥尔加
工业界高级人工智能发展中心的专家

举办组织 国际人工智能联盟



aianet.org

AI Alliance Network

一个全球性的合作组织

20+

国家

25+

人工智能协会

7000+

关联合作伙伴

