执行摘要

关于本报告

- 这是第一份《先进人工智能安全国际科学报告》的中期报告。由 75 位人工智能（AI）专家组成的多元化团队为本报告做出了贡献，其中包括由 30 个国家、欧洲联盟（EU）和联合国（UN）提名组成的国际专家咨询委员会。
- 在本报告主席的领导下，撰写本报告的独立专家们对报告内容拥有完全自主决定权。
- 在人工智能发展取得空前进展之际，本报告的首份出版内容将重点限制在近年来发展尤为迅速的一类人工智能——通用人工智能，即可以执行各种任务的人工智能。在快速发展的背景下，通用人工智能的研究正处于科学发现阶段，尚未形成科学定论。
- 只有对人工智能的风险进行适当管理，全世界的人们才能安全地享受通用人工智能的诸多潜在益处。本报告的重点是识别这些风险，对于用于评估和减轻风险的技术方法进行评价。报告的目的不在于对通用人工智能可能带来的所有社会影响，包括其许多潜在益处进行全面评估。
- 这份中期报告有史以来第一次汇集了由 30 个国家、欧盟和联合国提名的专家以及其他世界领先的专家，为有关通用人工智能安全的讨论和决策提供了一个共同的科学循证基础。我们在通用人工智能的能力、风险和风险缓解措施等方面仍然存在一些大大小小的分歧。但我们认为，这个项目对于增进我们对这项技术及其潜在风险的集体理解，朝着形成共识和有效风险缓解措施的方向更进一步，从而保证人们能够安全地享受通用人工智能的潜在益处，是至关重要的。事关重大。我们期待继续推进这项工作。

执行摘要要点

- 如果治理得当，通用人工智能可以用于促进公共利益，带来更高的福祉、更多的繁荣和新科学发现。然而，功能失常或恶意使用通用人工智能也可能造成伤害，例如在高风险环境中做出有偏见的决定，或通过诈骗、虚假媒体或侵犯隐私来造成伤害。
- 随着通用人工智能能力的不断进步，可能会出现大规模冲击劳动力市场、利用人工智能进行黑客攻击或生物攻击、社会失去对通用人工智能的控制等风险，尽管研究人员对这些情况发生的可能性存在争议。对这些风险的不同看法往往源于人们对社会为限制这些风险将采取的措施、这些措施的有效性以及通用人工智能能力的发展速度存在不同的预期。
- 未来通用人工智能能力的发展速度存在很大的不确定性。一些专家认为，到目前为止最有可能发生的是进展速度放缓，而另一些专家则认为，进展速度极快是可以发生或很有可能发生的情况。
- 对于评估和减少通用人工智能带来的风险，开发人员和监管机构目前可以采用的技术方法有很多，但它们都有局限性。例如，目前用于解释为什么通用人工智能模型会产生某种特定输出的原因的技术非常有限。

- 通用人工智能技术的未来是不确定的，即使在不久的将来也可能出现各种不同的轨迹，正面成果和负面成果都有可能。但是人工智能的未来并非不可避免。决定人工智能未来的将是社会和政府的决策。这份中期报告旨在促进关于这些决策的建设性讨论。

本报告综述了对于通用人工智能（可执行各种任务的人工智能）的科学认识的现状，重点在于理解和管理其风险。

使用人工智能的系统能力正在迅速提高，这突显了人工智能为商业、研究、政府和个人生活创造的众多机遇。同时，也提高了人们对与先进人工智能相关的当前危害和未来潜在风险的意识。

《先进人工智能安全国际科学报告》的目的是朝着在人工智能风险以及如何缓解风险方面形成国际理解共识的方向迈进一步。《报告》的首份中期报告将重点限制在其能力发展尤为迅速的一类人工智能——通用人工智能，即可以执行各种任务的人工智能。

在快速发展的背景下，通用人工智能的研究正处于科学发现阶段，尚未形成科学定论。本报告概述了对于通用人工智能及其风险的当前科学理解，包括确定具有科学共识的领域以及存在不同观点或研究问题未决的领域。

只有对通用人工智能的风险进行适当管理，世界各地的人们才能安全地享受其潜在的好处。本报告的重点是识别通用人工智能的风险，对评估和减轻风险的技术方法进行评价，包括有益地利用通用人工智能来减轻风险的方法。报告的目的不在于对通用人工智能可能带来的所有社会影响，包括其许多潜在益处进行全面评估。

根据多项指标显示，通用人工智能的能力近年来增长迅速，而对于如何预测未来进展却没有达成共识，因此各种情况似乎都有可能发生。

根据多项指标显示，通用人工智能的能力正在迅速发展。五年前，领先的通用人工智能语言模型极少生成一段连贯的文字段落。如今，一些通用人工智能模型可以就广泛的话题进行多轮对话，编写简短的计算机程序，或根据描述生成视频。然而，通用人工智能的能力很难进行可靠的估计和精确的定义。

通用人工智能的发展速度取决于技术进步的速度和监管环境。本报告侧重于技术方面，不讨论监管工作可能会如何影响通用人工智能的开发和部署速度。

近年来，人工智能开发人员主要通过不断增加用于训练新模型的资源（这一趋势被称为“规模化”）和改进现有算法，迅速提升了通用人工智能的能力。例如，最先进的人工智能模型用于训练的计算资源（“算力”）每年增长约 4 倍，训练数据集规模增长 2.5 倍，算法效率（相对于算力的性能）增长 1.5-3 倍。至于“规模化”是否在基础挑战（如因果推理）上取得了进展，研究人员之间还存在争议。

通用人工智能能力的未来发展速度对管理新兴风险具有重大影响，但即便面对不远的将来会发生的事情，专家之间也存在分歧。专家们对于通用人工智能能力缓慢、快速或极速发展的可能性持有不同程度的看法。这一分歧涉及一个关键问题：继续“规模化”资源和完善现有技术是否足以取得快速进展并解决可靠性和事实准确性等问题，还是需要新的研究突破才能大幅提升通用人工智能的能力？

几家开发通用人工智能的领先公司正押注“规模化”会继续带来性能上的提升。如果最近的趋势继续下去，到 2026 年底，一些通用人工智能模型的训练所使用的算力将是 2023 年发布的算力最大的模型的 40 倍到 100 倍，同时使用这些算力的训练方法的效率将提高 3 倍到 20 倍。然而，进一步增加数据和算力存在瓶颈，包括是否可以获取数据、人工智能芯片、资本开支和地方能源供应能力。开发通用人工智能的公司正在努力克服这些潜在瓶颈。

有几项研究工作旨在更可靠地理解和评价通用人工智能，但我们对通用人工智能模型和系统运作方式的整体理解还很有限。

管理通用人工智能风险的方法往往基于这样一种假设，即人工智能开发者和政策制定者可以评估通用人工智能模型和系统的能力和潜在影响。但是，尽管技术方法可以帮助进行评估，但所有现有方法都有局限性，不能对通用人工智能相关的大多数危害提供有力的保证。总体而言，对通用人工智能的内部运作、能力和社会影响的科学理解非常有限。专家们普遍认为，应优先提高我们对通用人工智能的理解。其中的一些关键挑战包括：

- 开发人员对通用人工智能模型的运行方式仍然知之甚少。这是因为通用人工智能模型并不是按照传统意义上的方式编程的。相反，它们是经过训练得来的：人工智能开发人员设定一个包含大量数据的训练过程，训练过程的结果就是通用人工智能模型。这些模型可能由

数万亿个称为参数的组件组成，它们大部分的内部工作原理对于包括模型开发人员在内人来说都是难以理解的。模型解释和可解释性技术可以提高研究人员和开发人员对通用人工智能模型运作方式的理解，但这方面的研究还处于起步阶段。

- 通用人工智能主要通过对模型或系统的各种输入进行测试来进行评估。这些抽查有助于评估其优缺点，包括漏洞和潜在的有害能力，但不能提供量化的安全保证。由于通用人工智能系统在不同的情况下、面对不同的用户或对其组件进行额外调整时可能会有不同的表现，因此这些测试往往会遗漏危险，高估或低估能力。
- 原则上，独立行动方可以对一家公司开发的通用人工智能模型或系统进行审计。然而，公司通常不会向独立审计人员提供必要的直接接触模型的权限，或进行严谨的评估所需的数据和所用方法的信息。一些国家的政府正在开始建设进行技术评价和审计的能力。
- 评估通用型人工智能系统对下游社会的影响很困难，因为风险评估研究还不足以产生严谨而全面的评估方法。此外，通用型人工智能具有广泛的应用场景，这些场景通常未预先定义，只是受到轻微的限制，从而使得风险评估更加复杂。要理解通用人工智能模型和系统的潜在下游社会影响，需要进行细致入微的多学科分析。在通用人工智能的开发和评估过程中增加不同观点的代表性是一项持续的技术和制度挑战。

通用人工智能会给个人和公众的安全与福祉带来严重风险。

本报告将通用人工智能风险分为三类：恶意使用风险、故障风险和系统性风险。此外，报告还讨论了导致许多风险的几个交叉因素。

*恶意使用。*与所有强大的技术一样，通用人工智能系统也可能被恶意使用，造成伤害。可能出现的恶意使用类型多种多样，既有相对证据充分的类型，如借助通用人工智能实施的诈骗，也有一些专家认为未来几年可能出现的类型，如恶意使用通用人工智能的科学能力。

- 通过通用人工智能生成的虚假内容对个人造成伤害的情况是通用人工智能恶意使用中相对有据可查的一类。通用人工智能可用于提高诈骗和欺诈的规模和复杂程度，例如通过通用人工智能强化的“网络钓鱼”攻击。通用人工智能还可用于在未经个人同意的情况下生成假冒这个人的虚假内容，例如未经同意的深度伪造色情内容。
- 另一个令人担忧的领域是恶意使用通用人工智能制造虚假信息和操纵公众舆论。通用人工智能和其他现代技术使得虚假信息的生成和传播变得更容易，包括试图影响政治进程。水印内容等技术反
- 通用人工智能也可能被恶意用于网络犯罪，提升个人的网络专业知识，使恶意用户更容易进行有效的网络攻击。通用人工智能系统可用于规模化和部分自动化某些类型的网络操作，如社会工程学攻击。不过，通用人工智能也可用于网络防御。总体而言，目前还没有任何实质性证据表明通用人工智能可以自动执行复杂的网络安全任务。

- 一些专家还担心，通用人工智能可能被用于支持武器的开发和恶意使用，如生物武器。目前还没有强有力的证据表明当前的通用人工智能系统会带来这种风险。例如，尽管目前的通用人工智能系统在生物学相关能力上展现出其能力不断增强，但现有的有限研究并未提供明确证据，证明目前的系统能使恶意行为者“加强”，让他们比通过互联网更容易获得生物病原体。然而，对未来大规模威胁的评估几乎很少，所以很难排除它不会发生的可能性。

故障风险。即使用户无意造成伤害，通用人工智能的故障也可能导致严重风险。这种故障可能有多种原因和后果：

- 用户可能对基于通用人工智能模型和系统的产品功能知之甚少，例如由于沟通不畅或误导性广告。如果用户以不适当的方式或出于不适当的目的部署这些系统，就会造成伤害。
- 人工智能系统中存在偏见是一个基本上有充分证据的问题。而对于通用人工智能来说，这个问题也仍未解决。在种族、性别、文化、年龄和残疾等受保护特征方面，通用人工智能的输出可能存在偏见。这可能会在包括医疗保健、工作招聘和金融借贷等高风险领域带来风险。此外，许多广泛使用的通用人工智能模型主要是在西方文化代表过多的数据上进行训练的，这可能会增加对未被这些数据充分代表的个人造成伤害的可能性。
- “失控”情景是未来可能出现的情景，在这种情景下，社会无法再对通用人工智能系统进行有意义的约束，即使这些系统显然正在造成危害。人们普遍认为，目前的通用人工智能缺乏构成这种风险的能力。一些专家认为，目前开发通用*自主*人工智能——能够自主行动、规划和追求目标的系统——的努力一旦成功，就可能导致失控。专家们对失控情景的可信度、可能发生的时间以及缓解这些情景的难度存在分歧。

系统性风险。通用人工智能技术的广泛开发和采用会带来一些系统性风险，从潜在的劳动力市场影响到隐私风险和环境影响：

- 通用人工智能，尤其是在其进一步快速发展的情况下，具有把非常多种多样的任务自动化的潜力，这可能会对劳动力市场产生重大影响。这可能意味着许多人会失去现有工作。不过，许多经济学家预计，潜在的工作岗位损失可能会被新创造出来的工作岗位和未被自动化的行业需求的增加所抵消，甚至完全抵消。
- 目前，通用人工智能的研发主要集中在少数几个西方国家和中国。这种“人工智能鸿沟”是多方面原因造成的，但部分原因是对开发通用人工智能所需的算力的获取水平不同。与高收入国家和技术公司相比，低收入国家和学术机构算力获取渠道较少，因此处于不利地位。
- 由此导致的通用人工智能开发市场集中化会使社会更容易受到若干系统性风险的影响。例如，在金融或医疗保健等关键领域广泛使用少数通用人工智能系统，可能会因为例如错误或漏洞等原因，导致这些相互依存的领域同时出现大面积故障和中断。

- 通用人工智能的开发和部署过程中日益增长的算力使用迅速增加了与通用人工智能相关的能源使用量。这一趋势没有缓和的迹象，可能会导致二氧化碳排放量和用水量进一步增加。
- 通用人工智能模型或系统可能会对隐私构成风险。例如，研究表明，通过使用对抗性输入，用户可以从模型中提取包含个人相关信息的训练数据。对于未来在敏感的个人数据（如健康或财务数据）上训练的模型来说，可能会导致特别严重的隐私泄露。
- 通用人工智能开发中的潜在版权侵权对传统知识产权法以及关于数据的同意、补偿和控制制度构成了挑战。不明确的版权制度使得通用人工智能开发者不愿公布他们使用的数据，也使得未经创作者同意就将其作品用于训练通用人工智能模型的创作者所能获得的保护不明确。

交叉风险因素。与通用人工智能相关的风险的基础是几个交叉风险因素，这些因素作为通用人工智能的特点增加了不是一种而是几种风险的可能性或严重性：

- 技术上的交叉风险因素包括：难以确保通用人工智能系统可靠地按照预期运行，我们对其内部运作缺乏理解，以及通用人工智能“智能体”的不断发展，智能体可以在减少监督的情况下自主行动。
- 贯穿各领域的社会风险因素包括：技术进步的速度与监管响应的速度之间可能存在差距，以及人工智能开发者因竞争激烈，可能会以牺牲全面的风险管理为代价而迅速发布产品。

虽然有几种技术方法可以帮助缓解风险，但目前已知的方法都不能够针对通用人工智能的相关危害提供有力保证或保障。

虽然这份报告没有讨论缓解通用人工智能风险的政策干预，但讨论了研究人员正在取得进展的缓解风险技术的方法。尽管取得了这些进展，当前的方法还不能可靠地预防现实世界中明显有害的通用人工智能输出。有几种技术方法可用于评估和缓解风险：

- 在训练通用人工智能模型使其更安全地运行方面取得了一些进展。开发人员还训练模型更能抵抗旨在使其失效的输入（“对抗训练”）。尽管如此，敌对者通常可以以较小到中等的努力找到替代输入，降低保卫措施的有效性。将通用人工智能系统的能力限制在特定使用场景中，有助于减少意外故障或恶意使用造成的风险。
- 有几种技术可以在通用人工智能系统部署后用于识别风险、检查系统操作和评价性能。这些做法通常被称为“监测”。
- 在通用人工智能系统的整个生命周期中，包括设计、培训、部署和使用，都可以减少系统中的偏差。然而，要完全预防通用人工智能系统中的偏见具有挑战性，因为这需要系统化的训练数据收集、持续评价和有效识别偏见。这可能还需要在公平性与准确性和隐私等其他目标之间进行权衡，决定哪些是有用的知识，哪些是不应反映在输出中的不良偏见。
- 隐私保护是一个活跃的研发领域。只要在训练中尽量减少使用敏感的个人数据，就能大大降低隐私风险。然而，当敏感数据被有意或无意地使用时，现有的降低隐私风险的技术工具很难扩大规模到大型通用人工智能模型，也无法为用户提供有意义的控制。

结论：通用人工智能可能有多种发展轨迹，这很大程度上取决于社会和政府如何采取行动。
通用人工智能技术的未来是不确定的，即使在不久的将来也可能出现各种不同的轨迹，正面成果和负面成果都有可能。但是人工智能的未来并非不可避免。通用人工智能如何发展、由谁发展；它的设计是为了解决哪些问题；社会是否能充分发挥通用人工智能的经济潜力；谁能从中受益；我们让自己暴露在什么风险类型之下；以及我们为缓解风险而进行的研究投入多少——这些问题和许多其他问题都取决于社会和政府今天和未来为塑造通用人工智能的发展而做出的选择。

为了促进对这些决定的建设性讨论，本报告提供了关于管理通用人工智能风险的科学研究和讨论的现状进行了概述。事关重大。我们期待着继续这项工作。