Google

以负责任的方式开发 AI 技术

简介

科学的进步更加凸显了全球问题带来的挑战，并且也加剧了问题的复杂程度。幸运的是，科学的进步也催生出了可以帮助我们解决这些问题的新型技术工具，其中最值得关注的就是 AI。目前，AI 研究不断取得突破，AI 在现实世界中的应用也在快速发展，这几乎为所有领域提高生产能力和洞察能力都带来了新的可能。与此同时，AI 也为我们解决旧有难题带来了新的曙光（有时也会放大问题）。例如，人类社会如何对待公平，如何打造包容的环境，以及如何让劳动者为未来的工作做好准备？

我们相信，如果运用得当，AI 可以为经济和社会带来巨大好处，并可以协助我们做出更公平、更安全、更全面、更明智的决策。但要实现这样的愿景，我们需要高度谨慎并不断努力，包括应对 AI 遭到滥用的风险，以及采取措施最大限度地降低此类风险。

AI 对世界的影响将取决于人们在利用这项技术时做出的种种选择。就像音乐家根据具体的听众、演出目的和场地限制来选择哪些乐器和音乐一样，程序员和企业也是如此，他们会在政府机构设定的界限以及文化接受程度内选择要应用哪些技术以及要实现什么结果。要如何利用 AI 带来的好处，以及如何构建有利于社会发展的框架，最终将由各个国家/地区和社会自行选择。Google 致力于在力所能及的范围内提供帮助，并希望本文档能够对关于 AI 话题的探讨起到一定帮助。

Google 的方法简介

作为 AI 领域的引领企业之一，我们深知 Google 有义务以审慎且负责任的方式开发和应用 AI，并为推行这样的做法做出贡献。和业界其他企业一样，我们列出了对我们至关重要的一般原则（见框中内容），并致力于制定相关流程和治理结构，以便在精神和现实层面恪守这些原则。

在实践中，AI 原则的实施方式常常取决于技术上的可行性。Google 处于 AI 研究的最前沿，致力于协助我们自身以及他人利用这项技术实现更多可能，包括面向开发者提供工具和指南（见下页框中的内容）。

Google 研究团队下设的人类与 AI 研究 (PAIR) 小组就是一个很好的例子。该小组致力于协助解决算法偏见、可解释性和可用性等方面的问题，他们开发了开放源代码工具，以协助开发者更好地了解其系统存在的风险，并帮助他们找出并解决可能存在的问题（例如 Facets 等数据集可视化工具）。概括来说，他们会发布研究成果，并在更大的范围内分享研究成果，以便让人们了解这项尖端的技术，并促进这项技术的发展。

Google 的 AI 原则

- **AI 应造福社会：**可能会给人类和社会带来的好处远远超出可预见的风险和弊端
- **AI 不应产生或加深偏见：**避免给特定人群造成不公平的影响，尤其是对于具备敏感特征（例如种族、民族、性别、国籍、收入、性取向、能力以及政治信仰或宗教信仰）的人群
- **AI 的开发和测试应以安全为宗旨：**根据 AI 安全性研究方面的最佳做法以适当的审慎态度进行设计，包括适当时在受限环境中进行测试并予以监控
- **AI 应对人负责：**提供适当的机会，让人们可以提出反馈、获取相关说明以及提交请求，并获得适当的人为指导和管理
- **AI 应纳入隐私保护设计原则：**鼓励采用隐私保护架构，并让用户了解和掌控数据的使用
- **AI 的开发应秉持严苛的科学卓越性标准：**要进行技术创新，必须采取科学的方法，并注重开放式调研、知识严谨性、完整性和协同合作
- **AI 应可用于符合上述原则的用途：**我们将努力限制可能有害或可能存在滥用情况的应用方式

Google 的部分工作重点

可解释性

通过开放相关工具的源代码并发布研究成果，Google 致力于提升 AI 的可解释性和可说明性。例如：

- **Tensor Flow Lattice:** 可让任何人训练灵活的模型，以便提前了解输入是否只能对输出产生正面影响
- **Tensor Flow 调试程序：**可让开发者在训练模型期间深入了解模型
- **可解释性的基石（Building blocks of interpretability）：**阐述了如何结合使用不同的技术来提供强大的接口，以便对神经网络输出进行解释

隐私保护

长期以来，Google 一直为研发适用于 AI 系统的隐私保护和匿名化技术提供支持，包括发布新的开放源代码（作为隐私保护最佳做法）。例如：

- **我们的开放源代码 RAPPOR 技术：**首个在全球范围内部署的大规模数据收集机制，可提供强有力的本地差分隐私保护
- **针对联合学习模型更新的安全聚合协议：**以加密方式为每位用户的更新提供强大的隐私保护（只对大批参与者的更新计算平均值）

安全性

目前，AI 系统面临的最大威胁之一来自“对抗攻击”，即不法分子通过对模型输入进行不易被人工检测到的细微更改来达到欺骗系统的目的。幸运的是，此类攻击很难实施，因此尚不普遍，不过，在应对此类攻击方面，Google 研究人员走在最前沿。已公开发布的研究成果包括：

- **对抗 Logit 配对 (ALP)：**抵御对抗样本方面最先进的技术
- **集成对抗训练：**在 ALP 推出之前，是抵御黑盒对抗样本方面最先进的技术，与斯坦福大学合作开发
- **CleverHans：**由 Google 团队维护的机器学习安全性研究库

公平性

训练数据遗漏和过度表示的内容会对 AI 系统造成影响。如果数据、模型设计、训练方法以及测试方法中存在人为偏见，则可能会导致得出的结果对不同人群产生不同影响。在机器学习公平性这个新兴的研究领域，解决这些有差别的结果是主要目标之一。Google 致力于为这一领域做出积极贡献，包括提供开发者工具。例如：

- **Facets:** 交互式可视化工具，可让开发者以不同的精细程度了解其训练数据的整体情况
- **Mmd-critic:** 探索性数据分析工具，用于查找数据中具有统计意义的少数样本

除了这些原则，我们还承诺不会涉足某些应用领域。具体而言，我们不会为以下应用领域设计或部署 AI：旨在造成或直接加剧人身伤害的武器或其他技术；或出于违反国际公认规范的监视目的收集或使用信息的技术；或目的与广泛接受的国际法和人权原则相悖的技术。 概括来说，对于任何存在重大的伤害风险的 AI 应用，只有在认定好处远远超出风险时，我们才会继续开展相关工作，并且会实施适当的安全限制。

我们会将这些原则和限制纳入到开发工作、实施审核以及商业协议中。我们将组织多个核心团队来审查隐私保护、歧视和法律合规等问题，并按产品领域分别进行评估。我们将成立跨职能的 AI 审核委员会来评估具有挑战性的问题。此外，我们还会将各种监管机制和工作惯例纳入到现有工作流程中。例如：

- 每个新产品或每项重要功能都要经过内部发布流程，流程中包括检查相应产品或功能是否遵循了主要原则。
- 2017 年，Google 的信任与安全团队试行了一项计划，目的是为产品团队提供专家协助，以便评估风险并测试是否存在偏见。现在这项计划已在整个公司范围内正式推行。相关支持包括提供模板、提示、案例研究、各种“dogfooding（内部测试）”测试组，以及实验设计方面的实践帮助。
- 对于采用 AI 的新工具，我们的隐私保护工作组和机器学习公平性团队将评估相关问题。
- 我们为每个产品领域指派了多名产品顾问，专门负责审核新产品的发布是否符合国际法律的规定。
- 我们正在试验各种机制，以便开发者更清楚地了解数据集的限制，进而更好地选择适合其应用的数据集。例如，一种可能性是添加一致的数据集相关信息（类似于食品包装上的营养成分标签）。出于类似原因，我们正在探索如何提供更详细的指南，以便使用者以适当方式使用我们面向开发者提供的预训练模型。
- 我们认识到，在产品开发的每个阶段都需要考虑道德问题。为了协助我们的团队做到这一点，我们正在试验各种方式，例如将相关模块（例如用于消除算法偏见的模块）添加到我们的内部机器学习训练中，以及与大学开展更多正式合作，以准备道德问题案例研究和自定义课程。

尽管我们努力让所有人都可以使用 AI（见框中内容），但需要注意的是，关于用户是否会将 Google 的 AI 工具用于我们无法预见或无法容忍的用途，我们在内部展开了激烈的辩论。

在考虑销售或分发我们可以预见到可能会被滥用的技术时，我们会考虑一系列因素，其中包括：相应技术是可以普遍获得，还是 Google 专有；相应技术能够被轻易改造以用于有害用途的程度，以及可能会造成多大影响；与我们的参与程度相关的风险可能性。

最后，我们认识到这不是 Google 能够或应该独自设法解决的问题。在讨论如何以负责任的方式开发和应用 AI 时，让广泛的相关方参与其中并广泛听取各方观点至关重要。为了促进这种讨论，我们会定期通过学术会议、行业会议以及政策论坛与外部专家沟通交流。Google 还与其他企业共同成立了 Partnership on AI，旨在研究并制定 AI 技术方面的最佳做法、加强公众对 AI 的了解，以及作为一个开放的平台，方便大家围绕 AI 及其对人类和社会的影响进行讨论和交流。

Google 致力于让每个人都能使用 AI

AI 能够蓬勃发展的部分原因在于有一系列鼓励公开发布和分享研究成果的公共规范。Google 致力于通过不断发布研究成果并积极参与相关会议，践行这些社区原则。我们还会发布开放源代码工具，以供研究人员和其他专家使用。例如：

- 我们对外发布了 Google 的内部机器学习工具包 TensorFlow 的源代码，让所有人都可以在此领域开展相关实验，并推动该技术向前发展
- 我们投资创建并分享了一些大型数据集，以针对多种数据类型为机器学习研究人员提供协助，这些数据集类型包括语音指令、照片和视频、在线讨论、音效和众包绘图
- “Learn With Google AI”网站为希望学习如何使用 AI 的人员（无论经验多寡）提供免费课程、教程和实践练习
- 最后，Google Cloud 降低了 AI 的入门门槛，让尽可能多的开发者、研究人员和企业都能使用 AI。我们的平台提供基于预训练模型构建的新型机器学习服务，可让商业应用实现无与伦比的规模和速度。

政策制定者如何提供帮助

和所有技术一样，AI 会产生哪些影响并不是注定的。尽管 AI 研究人员可以针对技术上可行的影响奠定坚实的基础，但 AI 实际上会产生哪些影响将取决于行业和社会对这项技术的需求，以及政府针对技术的应用提供的指导和设定的界限。因此，在规划愿景以及为 AI 发展构建支撑框架方面，政策制定者发挥着至关重要的作用。

尽管不同国家/地区的相关性和可行性有所差异，但是有一些共同的主题值得希望以负责任的方式使用 AI 的政策制定者去探讨。让普通公众和企业对 AI 放心并支持和鼓励他们使用 AI 至关重要，清楚阐明治理框架也非常重要。让政府在以负责任的方式使用 AI 方面树立榜样会直接带来很多好处，还可以为社会提供最佳做法示范。最后，与 AI 相关的某些挑战是由数据访问和研究经费导致的，而政策制定者恰恰能够协助克服这些挑战。

下面，我们对其中的各个主题进行了更详细的说明，并根据世界各地的示例提供了一些能够带来启发的应对策略建议。我们希望这有助于激发人们在政策方面的想法，并期待各方更加踊跃地参与到 AI 的发展中来。

鼓励以负责任的方式使用 AI 的行动方针

1. 协助提升公众对 AI 的信心和了解
2. 鼓励在重点行业领域优先采用 AI 技术
3. 为 AI 研究提供便利，协助克服实施障碍
4. 鼓励以负责任的方式分享数据，以便为训练 AI 系统提供更多数据
5. 推动制定具有建设性的治理框架，并帮助治理机构人员积累专业知识
6. 让政府成为以负责任的方式使用 AI 的榜样
7. 采取措施，为劳动力转型做好准备

1. 协助提升公众对 AI 的信心和了解

AI 为造福社会创造了绝佳机会，可在推动科学发展、让人们更轻松地享受医疗服务以及提高经济生产率方面提供重要帮助。不过，要实现这些愿景，需要社会接受这项技术。目前，科幻小说中虚构的情节影响了公众对 AI 的信心，此类情节往往会轻松占据头条位置，导致人们忽略那些虽然不太引人注意但更加迫切的危险（例如偏见和恶意使用等风险）。人们对 AI 的恐惧还常常掺杂着对未来工作形态和日益加剧的不平等的普遍担忧。承认并消除此类担忧，以公平现实的方式向不同专业和不同社会阶层的人群介绍由此带来的机遇和挑战，需要政府和业界共同努力，政府部门更是在其中发挥着关键作用。在公众对 AI 技术的信任程度非常低的地区，还可以开设专家咨询论坛，提供相关框架来推动实现更正式的持续互动和监控，借此打消公众疑虑，让人们信任 AI。

可以考虑开展的活动

- 开设论坛，以收集并整合社会各界的观点
- 举办面向公众开放的专家讲座，为针对重要主题开展更加有理有据的讨论提供便利条件
- 组织一系列公民会审，针对重要的 AI 问题展开讨论并提供建议
- 开展正式的调查，收集意见并表明对公众意见的重视
- 成立专家咨询委员会，为公众参与提供渠道
- 寻求国家科研机构和非政府组织的支持，以便开展富有创意的公众参与活动，例如知名科学家讲座、跨界艺术/科学活动
- 开展培训活动，重点介绍 AI 在日常生活中的应用（例如在公共部门服务、医疗保健等领域的应用），以吸引普通民众（而不仅仅是专家）重视这项技术，并让 AI 走进他们的生活
- 提供培训补助金，鼓励不同背景的人员学习 AI 知识，以便为 AI 发展注入新鲜观点，并实现更广泛的民众参与

2. 鼓励在重点行业领域优先采用 AI 技术

仅当业界以有意义的方式应用 AI 时，AI 的经济前景才能变为现实。这就需要透彻了解 AI 善于解决哪些类型的问题、当前局限或固有局限，以及实施 AI 解决方案所需的资源（工具、数据、专业知识、计算能力）。政府可以充当引路者，带动企业探索 AI 商机并进行投资。另一个非常重要的手段是政府提供大力支持，协助相关方针对应用的 AI 提供培训以及传播相应的最佳做法和标准。

可以考虑开展的活动

- 开展研究/问卷调查，以评估业界对 AI 的认知度、对使用 AI 的兴趣以及采用 AI 技术面临的障碍
- 指派专家组向企业提供有关优先事项的建议并宣传 AI，甚至可以提供类似于 AI 孵化器的创业资本/专家资源
- 针对利用 AI 技术的企业提供激励措施，激发他们对 AI 的兴趣
- 提供补贴（如电费折扣、更快速的资本支出折旧），支持在 AI 实体基础设施不足的地区投资基础设施建设
- 提供资金，支持在具有巨大社会需求的领域（如灾难应对）应用 AI
- 鼓励研究人员和企业在遵循隐私保护规范的同时，创建并分享与重要领域相关的数据集
- 寻找切实可行的方法（例如，提供更灵活的数据本地化规则），让企业可以更轻松地获取 AI 工具，包括通过基于云技术的服务获取 AI 工具
- 提供激励措施，推动计算机科学和其他行业重点学术领域之间（例如计算机科学和农业、计算机科学和医学）开展更多跨学科协作
- 鼓励大学在课程（而不仅仅是工程专业的课程）中增加 AI 应用方面的培训，让新一代毕业生在进入相应行业之前便已做好充分准备
- 资助开发和提供 AI 方面的职业培训，以便重点领域的从业人员或有志于在此类领域就业的人员能够接受这些培训

3. 为 AI 研究提供便利，协助克服实施障碍

尽管 AI 研究近来取得了一些进展，但仍面临许多挑战，只有在克服了这些挑战之后，AI 才能释放全部潜力。例如，AI 系统需要具有更好的可解释性，能够更加高效地利用大量数据和计算资源来训练模型，并可以让更多人更轻松地使用和构建模型。要实现AI 生态系统的蓬勃发展，不能只依赖私营部门推动的基础性的基本研究；对于新的高风险、高回报研究领域的研究，可能只有获得政府资助才能开展。此外，有些应用情况可能会带来更广泛的社会效益，学术机构以及由政府资助的研究组织则是对此类应用情况开展研究的最佳选择。

可以考虑开展的活动

- 为国家机构的 AI 研究和高级研究提供专项资金
- 在各地成立 AI 研究和应用方面的卓越研究中心
- 公开由政府资助的研究以及由此获得的数据集
- 在政府设立的实验室和研究机构围绕对国家具有重要意义的关键领域开展 AI 研究
- 制定框架，促进公共部门/私营部门合作开展 AI 研究

4. 鼓励以负责任的方式分享数据，以便为训练 AI 系统提供更多数据

机器学习模型需要数据集来进行训练，而精心挑选和定制的数据集有助于最好地体现和解决需要处理的问题。政府可以制定框架，鼓励人们创建、分享和重复使用与重点应用领域相关的数据集，并此创造便利条件，同时确保达到用户对保护隐私的期望，从而帮助推动开展 AI 研发工作。与此同时，对于很难在社会效益和个人权利之间做出取舍的情况，建议以更加清楚的方式对数据条例做出实用性解读。（例如，在欧洲境内，按照 GDPR 的规定，哪些用途算作科学用途？在欧洲境外，是否存在任何可以不优先遵循数据最小化原则的情况？）

可以考虑开展的活动

- 制定标准条款并建立标准机制，推动以有助于保护隐私的方式分享数据，进而帮助减轻协商此类交易的法律和管理负担
- 提供更多公开数据集，尤其是重点主题领域的数据集，以实现创新
- 为接受政府资助的研究人员提供激励措施，鼓励他们在遵守隐私保护规范的情况下，以机器可读的格式发布与其研究相关的数据集
- 寻找可让人们更轻松地导出和分享个人数据的方式，方便他们贡献个人数据以用于研究和其他用途
- 成立专家机构，以及时为研究人员提供意见和建议，并指导他们在社会效益和个人权利之间权衡利弊

5. 推动制定具有建设性的治理框架，并帮助监管机构积累专业知识

AI 可以通过多种方式影响人类社会，因此为了确保 AI 可以造福人类，政府机构可以和业界共同努力，并发挥重要作用。现在，已有一些行业（医疗保健、交通运输、通信，等等）法规对 AI 的实施做了相关规定。要评估 AI 在具体情况下的使用，以及评估新技术的影响和实施结果，行业专家通常是最佳人选，但他们可能需要相关方的支持，以掌握 AI 专业知识。随着 AI 的进步，政府机构应不断拓展自身的技术专业知识，并探索各种合作框架，以便最大限度地减少问题并发挥 AI 的潜能。此外，如果有根据共识制定的最佳做法以及自律机构，还将有助于开发出灵活、精细的方法。

可以考虑开展的活动

- 逐个领域进行分析，了解现有监管方案适用于 AI 系统的情况，以及有哪些不足之处（如有）
- 找出会妨碍人们以负责任的方式使用 AI 的既有限制，并寻求解决方案
- 示例：要检查相应系统是否存在种族偏见，必须先推断种族，但现有的反歧视法律和隐私保护法律可能会给推断种族造成障碍
- 示例：设备中采用的 AI 与云端 AI 具有不同的风险和特征，而“一视同仁”的数据保护规则可能无法反映此类细微差别
- 示例：版权规则可能会限制可用于训练 AI 系统的数据，但如果排除关键部分的数据，则可能会影响为减少偏见所做的努力
- 在最有可能遭到破坏的领域，提供资金来强化行业内监管者的内部技术专业知识
- 指派咨询委员会或主要联系人来协调 AI 治理问题，委员会成员包括 AI 研究社区代表、行业代表和民间团体代表
- 与世界各地的治理机构互动交流，分享经验并从经验中学习
- 鼓励行业分享最佳做法并宣传行为准则
- 对政府资助的研究人员进行道德规范培训（类似由美国国立卫生研究院资助的生物科学研究人员需要接受的研究道德规范培训）

6. 让政府成为以负责任的方式使用 AI 的榜样

无论是对于公共部门还是对于企业，AI 在提高生产能力和服务质量方面都有着巨大潜力。因此，政府可以率先示范最佳做法，在以下方面做出表率：如何优先把握各种机会以充分利用 AI，以及如何以切实可行的审慎方式应用该技术。更笼统地说，政府机构可以提供公开数据集以供其他人用于开发服务，因此在推动该技术不断进步方面，政府机构还发挥着重要作用（就像 Imagenet 数据库助力计算机视觉技术取得进步一样）。

可以考虑开展的活动

- 指派专家组针对 AI 的应用提供切实可行的指南
- 制定基于各项原则的基本指南，以便政府机构在资助包含 AI 组件的项目或采购相关系统时参考。将该指南纳入到针对相关活动的提案请求和监管工作中
- 鼓励开展试点计划，以加快 AI 在改善便民服务方面的应用
- 聘请专家或与专家合作，寻找各种机会来提升内部人员的 AI 专业知识
- 与 Kaggle 等平台合作，开展各种利用公开数据集协助改善公共服务质量的竞赛
- 确定面临的障碍，投资建设基础架构以及开展培训，鼓励政府机构完善新的公开数据集的整理和分享机制

7. 采取措施，为劳动力转型做好准备

人们普遍认为，AI 将会颠覆就业格局，但影响的步伐有多快、波及面有多广还是未知数。在帮助人们掌握未来所需技能方面，政府机构可以发挥关键作用，例如在学校中推行旨在提升数字知识水平的计划，或提供激励措施来促进在职学习。要确保这些技能得到善用，提高劳动力市场的灵活性和流动性至关重要，这样有助于企业更轻松地招聘优秀人才，而且有助于消除就业壁垒（例如过度依赖证书或认证）。与此同时，鼓励发扬创业精神和培养适应能力的计划有助于人们做好准备，及时发现并抓住机遇。最后，务必要反思社会安全保障网，并考虑是否需要根据不断变化的就业环境调整资助方案和相关规定，让社会安全保障网与时俱进。

可以考虑开展的活动

- 与重点行业开展合作，制定下一代学徒培训方案，尤其要为面临转型且经验丰富的员工制定培训方案
- 与雇主和员工代表保持联络，以便审核职业许可制度，并确保相关限制仍有合理的依据（例如出于公共安全考虑）
- 成立可以根据当地就业趋势协助设定职业培训优先事项的区域性专家组（包括学术代表、企业代表和劳工代表）
- 为愿意去异地工作的员工（及其家属）提供支持和激励措施，协助重点行业和地区的企业招聘到优秀人才
- 为所有成年人提供助学金和税收优惠储蓄账户，鼓励他们终身学习
- 为研究提供资助并推行试点计划，协助确定并推广在职培训的最佳做法，包括尝试使用 AI 来定制和提供及时教育
- 让求职者可以通过更多途径获取非正式（但经过证实）学习机会，例如有针对性的认证计划或技术“（集中）训练营”
- 研究并实验社会安全保障网方面的新方法

Google