牛津人工智能
与善治委员会

OXFORD
INTERNET
INSTITUTE

UNIVERSITY OF
OXFORD

公共服务中的人工智能：从原则到实践

牛津人工智能与善治委员会

2021 年 12 月

公共服务中的人工智能：从原则到实践

牛津人工智能与善治委员会

2021 年 12 月

牛津人工智能
与善治委员会

目录

报告摘要

早期经验表明，随着世界各国政府在公共服务中应用人工智能，这会带来严峻挑战。

然而，这也是一个机遇，通过合作，人工智能系统将有助于实现善治，并帮助我们解决一些最紧迫和最棘手的公共问题。

牛津人工智能与善治委员会研究和审议了确保人工智能系统有效用于公共服务的主要挑战，并提出了三个关键*问题*：

- 谁应该为公共服务中的人工智能提供指导？
- 我们如何为人工智能建设公共服务能力，以实现善治？
- 我们如何确保公共服务中的人工智能是值得信赖、可信赖的？

我们确认，为了使用人工智能实现公共利益，必须满足*几个目标*：

- 人工智能的设计必须具有包容性。
- 人工智能的任何采购必须由知情的公共机构指导。
- 在公共服务中实施人工智能必须有目的性。
- 人工智能系统必须始终对利益相关者负责。

为了实现这些目标，我们提出了三项*建议*：

1. 在国际层面，政府、工业界和民间社会必须共同努力，建立和赋能 (a) 一个国际**科学机构**来推进公共服务人工智能应用的研究；以及 (b) 一个**仲裁机构**，以裁决公共服务人工智能系统中涉及的利益相关者之间可能出现的争议。
2. 在这两个新的国际组织的支持下，各国政府必须 (a) **建设其公共服务的能力**，以深入参与公共服务人工智能系统的设计、采购、实施和当责；以及 (b) **为公共机构的工作人员提供工具箱**，以完成其监督工作。
3. 必须通过关于日常应用和即将出现的实际用例、其影响和风险的公共教育活动来加强人们对人工智能使用的信任，方法是 (a) 要求政府**公开披露**人工智能技术在公共服务中的使用情况；以及 (b) 引入一个多部门机构，由其**提供一个基本的认证**系统，不断验证应用程序，以随着时间的推移建立信任。

以下三项直接行动有助于这些目标的实现：

I. 为科学机构和仲裁机构开展可行性研究。

II. 就这两个机构如何扩展和补充现有能力和最佳实践，与现有国家机构和多边机构进行沟通协商。

III. 计划在未来 36 个月内参与国际日程上已经排定的几个重要的技术创新里程碑活动。

前言

人工智能有望解决公共政策和社会变革中一些最紧迫的挑战，但我们的政府可能还没有准备好将其用于善治。

近年来，许多政府都试图利用机器学习、大数据和其他算法工具等新技术来为民众制定和实施更好的政策和计划。尽管公共部门的创新是值得鼓励的，但其中许多项目都难以顺利启动，与工业界的合作冰消瓦解，劣质数据和有缺陷的技术导致了往往带有偏见的意外结果。

这些经历引发了一个非常重要的问题：如何才能最好地将人工智能用于公共服务？

2020 年到 2021 年，为了专门回答这个问题，我们参加了一个由牛津大学支持的项目，即牛津人工智能与善治委员会 (OxCAIGG)。我们的团队成员包括来自世界各地的、担任各种职位、代表各方利益的独立的董事、高管、学者、律师和政府顾问。在 18 个月的时间里，OxCAIGG 举办了专家简报会，并与牛津大学的技术顾问进行了磋商，对与人工智能和善治相关的主题进行了原创性研究，并召开了几天的研讨会。

根据我们作为委员的观点、迄今为止的研究以及我们自己的经验，我们认为现在是独立的国际机构承担责任和使命的时候了，这些机构应当负责设计和支持实施标准和最佳实践，以便有效地在公共部门应用人工智能。

随着各国政府在采用这些技术方面取得进展，一些机构也同时起到了促进作用，这些机构可以提供有关工业界参与的指导，并提供避免有偏见或低劣的政策和方案结果的方法，还可以提供裁决或争端解决服务。这些机构也将使世界上许多在采用技术方面的制度或监管能力有限的政府受益。

作为一个委员会，我们编写了一系列研究报告，评估在公共服务中使用人工智能的机会和风险，并将这一经验应用到我们建议的制定中。我们的目标是减少使用人工智能的负面结果，并支持对这一新兴技术进行更标准化的全球治理。

我们的目标还包括提出结构性的建议，这些建议具有普适性，并且基于与我们的研究一致的一系列原则。这一办法将考虑到相关民众的反馈意见，为其灵活和创新的实施提供空间。

作为一个委员会，我们希望不仅局限于制定人工智能使用原则——这项工作一直在稳步进行——我们还希望制定具体的政策指导方针和组织框架，供相关国家和国际参与者考虑。

这份名为《公共服务中的人工智能》的报告并不是对相关研究文献的直接引述、综述——我们会在其他报告中列出我们所参考的相关资料。这是一份关于我们的审议和磋商的简要声明，旨在总结我们认为的关键问题和重要目标、我们的建议以及将人工智能用于善治的后续工作。

牛津人工智能与善治委员会委员

1导言

早期的经验和结果表明，随着世界各国政府在公共服务中应用人工智能，这会带来严峻挑战。然而，这些早期的经验也给我们指明了方向——合作和共识标准可以在缓解挑战和充分利用机会方面产生重大影响。

牛津人工智能与善治委员会 (OxCAIGG) 调查了人工智能在公共服务领域面临的几个具体挑战，特别关注世界各地民主国家在公共项目中使用人工智能工具的情况。委员会这份最后报告的目标是评估挑战，并借鉴利用人工智能实现善治的核心原则，为克服这些挑战提出具体建议。

从 2020 年成立到 2021 年编制本总结报告，牛津人工智能与善治委员会：

- 资助并发表了七份原创性研究报告章和两篇关于人工智能和善治的当代问题的社论。
- 与来自工业界、民间社会和公共部门的利益相关者举行了八次专家电话会议。
- 与牛津人工智能与善治委员会的委员和技术顾问举行了四次圆桌磋商。
- 向国家、超国家和多边公共机构的高级决策者介绍了我们的工作。
- 应国家和多边公共机构的专家证词要求递交了证据。

这项原创研究，加上我们的专家简报和与政府官员的接触，为我们的审议和本报告提出的建议提供了重要的参考信息。在本报告中，我们概述了关键挑战，提出了一套基本原则，并确定了现在应该采取的行动。

提议的行动被总结为可操作的、基于证据的建议，这些建议将使政府和公共行政部门能够利用人工智能带来的好处。遵循 OxCAIGG 的基本原则，我们承诺：

- 以灵活且有目的性的方式行动，了解目前用于治理、公共管理、确保社会福利以及提供公共产品和服务的人工智能工具的影响。
- 协助政策制定者和企业家解决政策问题和公共应用，并提出将人工智能和机器学习应用于公共服务的理念。
- 确定解决方案并指导决策过程，以加强开发，引入人工智能来解决而不是使社会问题复杂化，并建立公众对使用这些工具的信任。

2 挑战

公共部门采用人工智能工具将人工智能技术固有的系统性风险与创新公共部门规划和决策相关的实际问题结合起来。

在应用这些工具时，政府将被迫改变他们采购、制定和实施计划的方式，并将遇到有关培训数据收集、复杂技术系统评估、工作人员培训和雇用新专家工作人员等问题。

OxCAIGG 开展的研究回顾了当今人工智能在治理和公共服务中最普遍的许多应用，并展望了未来，预测了将在不久的将来面临的一些挑战。我们的研究人员指出了阻碍采用公共服务人工智能的一系列挑战。我们的建议旨在缓解当前的挑战以及灵活适应未来挑战。在本报告中，我们对 OxCAIGG 开展的研究进行总结，然后讨论了这项工作带来的主要挑战。

全球对人工智能、机器学习和自动化决策的态度[1]

Lisa-Maria Neudert、Aleksi Knuutila 和 Philip Howard 在他们的研究报告中基于来自 142 个国家或地区的 154,195 名受访者样本的调查数据，分析了公众对人工智能参与个人事务和公共生活的潜在危害和机会的认知的基本指标。他们发现，不同地区和社会经济群体的公众对人工智能的认知差异很大。东西方存在显著差异，公众对人工智能的担忧程度在欧洲 (43%)、拉丁美洲 (49%) 和北美 (47%) 最高，而在东南亚 (25%) 和东亚 (11 %) 认为人工智能有害的人比例相对较低。在不同的职业中，企业高管和政府管理人员 (47%) 以及其他专业人士 (44%) 对人工智能的热情最高，而制造业工人 (35%) 和服务业工人 (35%) 则没那么感兴趣。

政府人工智能项目的实践经验：四个智慧城市计划的成果[2]

在本报告中，Godofredo Ramizo Jr. 调查了政府如何利用人工智能来提供公共服务——重点是人工智能驱动的智慧城市项目。通过广泛的文献回顾以及对香港、马来西亚和新加坡参与智慧城市和类似的人工智能驱动项目的高级政府官员的第一手采访资料，Ramizo 展示了政府人工智能项目的多样性，并确定了有助于维护公共利益的实用原则。研究报告显示，政府正在努力应对人工智能采购、实施和影响评估，并努力确定项目的财务、技术和政治可行性。尤其是在科技公司掌控优越的资源和影响力的情况下，政府的谈判地位就会受到挑战。

地方政府与人工智能[3]

Thomas Vogl 在这份研究报中探讨了英国地方当局对人工智能的使用情况。虽然有许多与后台自动化、决策支持的预测分析以及使用聊天机器人与民众互动相关的成功项目，但 Vogl 发现，政府在成功采用人工智能方面面临着重大的现实挑战。他表示，地方当局需要提高其数据收集和分析能力，在寻求人工智能解决方案之前更清楚地定义问题，并为供应商提供有关地方当局及其流程的背景知识的见解。

老问题，新技术：高度分裂和社会化分层的社会中的人工智能、人权和善治。肯尼亚的案例[4]

Nanjala Nyabola 在报告中调查了肯尼亚政府关于人工智能和区块链技术的政策，并评估了其成功情况。通过对政策文件和研究报告进行文献回顾和分析，她指出，在肯尼亚，这项技术的主要应用集中在平价医疗、食品安全、制造业、住房、网络安全和土地所有权方面。Nyabola 发现，在肯尼亚等高度分裂的社会中部署人工智能可能会加深现有的分歧，包括围绕阶级和身份的分歧，并且在工业环境和面向公众的环境中使用人工智能的伦理在其含义和社会影响方面有所不同。与许多发展中国家一样，肯尼亚现在才开始制定管理技术使用的法律框架。

监控即服务：欧洲人工智能辅助监控系统的大规模市场[5]

Yung Au 在她的研究报告中研究了生产人工智能辅助监控系统并将其出口到世界各国政府的欧洲市场。她研究了她所谓的“监控即服务”，包括为监控提供的服务和软件，这些服务和软件由复杂系统组成，这些系统具有用户友好的界面以及持续维护、更新和故障排除支持。她的分析集中在最近作为重大监管目标出现的这类服务的三个例子上：面部识别和分析；语音识别和分析；以及行为分析和同化系统。随着人工智能技术和大规模监控应用之间的重叠不断增加，潜在的危害也在增加。如果监管不力，这一市场将很有可能产生广泛的持久危害。

人工智能的协调：标准在欧盟人工智能法规中的作用[6]

在这份研究报告中，Mark McFadden、Kate Jones、Emily Taylor 和 Georgia Osborn 调查了技术标准在欧盟人工智能法规草案中规定的人工智能安全、公平和创新发展中的作用。研究报告显示，这种情况下的标准化是复杂的，标准与欧盟委员会目标之间的关系是利益相关者、经济利益和既定标准开发组织之间具有挑战性的交叉点。在广泛研究和与利益相关者协商的基础上，该法规草案为人工智能治理和标准制定了一个综合框架。该报告重点介绍了法规草案赋予人工智能标准的作用。具体而言，符合统一标准将为高风险 AI 应用程序和服务带来符合性推定，从而使人们对其符合拟议法规的繁重和复杂要求有一定程度的信心，并为工业界遵守欧洲标准创造强有力的激励。

关键挑战

这些报告和简报文件展示了原创研究，或以原创方式展示了一系列前沿研究。他们从越来越多的关于人工智能如何用于公共服务的社会和政策科学中汲取经验，并从世界各地的经验和案例研究中进行选择。它们确定了将人工智能用于公共利益领域的关键挑战。积极和建设性的案例带来启发，成效不佳的案例给我们警醒，并为未来的决策指明方向。

首先，一些研究强调了制定人工智能标准的制度和结构复杂性。虽然某种程度的标准化有明显的好处，但需要有科学支持的标准化路线图和人工智能标准化的专注能力。目前，没有有效的争端解决机制。

其次，在人工智能的采购以及培训数据的收集和分析方面，将人工智能用于公共服务存在非常现实的实际挑战。公务员缺乏专业知识和技能，也缺乏做出正确决策的实用工具箱。很明显，与政府和公共行政管理人员相比，强大的科技公司具有更强的议价能力和专业知识。公务员需要技术和实践能力才能采用人工智能促进善治。

第三，公共部门人工智能的使用所面临的核心挑战是公众的信任本身。公众对政府服务的信任始终至关重要，任何认为机器学习应用成本高昂、破坏平等或导致新问题的观念都会产生不必要的障碍。

最后，即使政府出于善意使用人工智能，也会加剧现有的偏见和不平等。公务员缺乏法律或实际框架，这对采用这些新技术构成了挑战。用于治理的人工智能技术可能会对人权产生系统性偏见、不可预见的后果，甚至系统性风险。监管不力的市场可以通过维持这些偏见和不平等来创造影响公共生活的事实上的技术标准，因为这些市场中的公共利益干预无法捕捉和纠正需要纠正的偏见和不平等。

联合研究

根据本研究中揭示的挑战，并根据我们的专家呼吁和内部圆桌讨论，我们确定了公共服务人工智能监管的三大核心不足。

谁应该为公共服务中的人工智能提供指导？绝大多数政府都认识到了人工智能的潜力，并希望将其用于善治。但人工智能技术的采用带来了非常具体的实际和规范性挑战，而政府本身并没有能力解决这些挑战。

我们如何为人工智能建设公共服务能力，以实现善治？公务员在公共服务中采用人工智能方面发挥着至关重要的作用。人工智能的明智实施所需要的一整套技术技能和教育是公共机构很少能获得的。

我们如何确保公共服务中的人工智能是值得信赖、可信赖的？我们的研究表明，公众对人工智能的信任是有争议的。我们必须确保人工智能系统是值得信赖、可信赖的，才能实现成功应用。

3 原则

今天的人工智能领域充满了道德规范和规范指南。世界领先机构的世界领先专家提出了在刑事司法、医疗和可持续发展等领域使用人工智能的原则。

OxCAIGG 无意增加给人工智能进一步施加这些来自“高层”的压力。相反，我们致力于借助团队的经验和专业知识，制定一套简短而明确的可操作建议，为公共部门使用人工智能指明方向。从委员会成立之日起，我们就希望明确表达我们的使命，即推动人工智能在公共服务中的具体应用，并促进善治。我们的原则可作为对框架的补充，这些框架足够灵活，具有全球普适性，并且足够精确，可以为公务员和政府官员在应用这些新工具时采取的非常具体的行动指明方向。[7]

OxCAIGG 的工作遵循以下四项关键原则，这些原则构成了我们作为委员会工作的基础以及本文件中提出的建议：

I **人工智能的设计必须具有包容性：**
人工智能工具和方案必须以公共部门专家的经验为依据，克服与使用不充分的数据集、排斥少数群体和代表性不足的群体以及设计缺乏多样性有关的歧视和偏见挑战。

II **人工智能的采购必须透明：**
这将克服人工智能工具的获取和开发、设计和可用性方面的挑战。采购流程必须包括对实施人工智能工具的风险和收益的评估。

III **人工智能的实施必须明智：**
公务员需要在互操作性、可解释性、偏见以及与决策过程的整合等问题上接受培训。

IV **人工智能必须当责：**
人工智能系统做出的决策必须透明，避免出现“暗箱”操作。包括引入监控和审计人工智能系统的流程。

4 建议

谁应该为公共服务中的人工智能提供指导？

委员会讨论了有关公共部门开发、采购和使用人工智能的基本问题。通过我们的讨论，我们确定了两个监管需求。首先是成立一个专门的国际科学机构，推进算法审计、社会影响、用例和最佳政策实践的研究，并推广此类研究，以激励和协调使用新的人工智能系统来解决需要集体行动的问题。其次是成立一个仲裁机构，以快速有效地解决公共使用人工智能系统的开发者、监管者和主体之间的纠纷。这两个组织将是独立但互补的：科学机构将向仲裁机构提供公正的证据；仲裁机构将表明有必要对提交给它的问题进行研究。

有关公共影响的科学咨询机构

第一个全球性机构是一个*科学咨询机构*，将致力于就人工智能、机器学习和其他先进算法系统对公共问题的影响进行科学、工程和技术对话。以政府间气候变化专门委员会 (IPCC) 为模式，该机构将协调科学对话，促进对技术发展和审计系统研究的同行评审，并确定关于以下方面的共识点：设计过程的包容程度如何、采购过程的有效程度如何、实施系统是如何工作的、人工智能系统对其服务的公众有何影响。不过，最重要的是，该机构应起到技术官僚和政策指导的角色。其领导团队必须能够为技术和工程决策的经济、文化和政治结果提供依据。

该科学机构应具备了解人工智能和社会系统如何相互作用的技术专长。这是一个天然具备多学科性质的项目，旨在解释人工智能如何对世界各地的文化、经济和政治生活产生具体影响，并收集和评估人工智能对个人人权影响的证据。实施包容性设计、知情采购、有目的的实施和持续的当责将需要计算机和社会科学家的关注。

对于世界上没有本国专家工程师和社会研究人员社区的许多政府来说，这将是一个重要的机构。一个配备适当人员的研究机构将负责评估成员国采取的政策路线，主持关于技术发展过程中

出现的挑战的高级别对话，评估社会平等可能如何变化的证据，并提供一个论坛，从工业界、民间社会以及和受公共服务领域人工智能影响的无数其他公共利益团体的见解和意见。

至关重要的是，这个科学机构将为政策制定者的标准和认证体系以及工具箱的形成提供证据。审计算法、解释影响和评估最佳实践的工作需要由计算机、社会和政策研究人员领导，才能具有可信度和影响力。

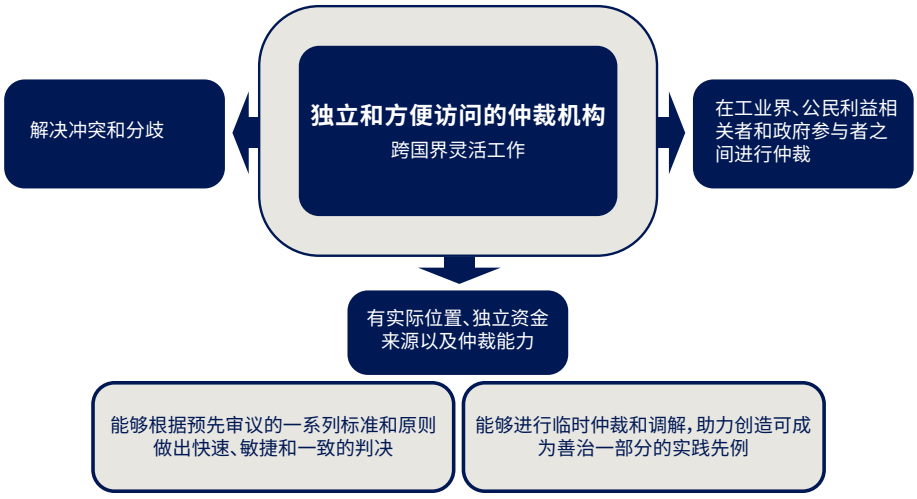
有几种国际学术协会、多边研究机构和政府间机构已经要求研究单位开始研究其中一些问题。然而，围绕公共服务人工智能全球影响的证据建立协调的科学共识并没有系统地组织起来。好在有 IPCC 为如何协调科学学习提供了范例。当然也有关于社会调查的主要领域如何随着时间的推移而僵化的研究，但我们对鼓励社会调查中的多样性和创造力以减轻这种情况有足够的了解。此外，这样一个科学机构的反馈意见对于今后的发展至关重要：我们在本报告建议的其他机构和流程将需要可靠的证据和科学共识才能加以使用。

独立和方便访问的仲裁机构

第二个机构是一*个全球仲裁机构*，将帮助解决冲突和分歧，在出现关键问题时在工业界、公民利益相关者和政府参与者之间进行仲裁。作为一个委员会，我们发现，创新速度快到令人难以置信。想要预测新的人工智能工具将如何在公共生活中应用很难。凭借其市场份额，工业界可以制定事实上的技术标准，而立法则滞后。我们有理由相信，分歧和误用将会发生，技术错误和设计缺陷都会使机器学习在公共生活中的应用变得复杂。从形式上讲，一些国家/地区有可以进行民事诉讼的合法途径。但许多国家/地区没有，许多这样的分歧跨越国界产生，而且绝大多数国家仲裁法院评估争端的技术能力有限。更重要的是，事实证明，创新的速度甚至超过了本应适用未来形势的立法。

一个设立了高效处理各类事项的秘书处的专门仲裁机构——具有实际位置、独立资金来源以及提供仲裁和调解服务的能力——将助力创造可成为善治一部分的实践先例。仲裁庭的任务是根据科学机构可能提出的一系列标准和原则做出快速、敏捷和一致的判决。当今技术发展的速度使得依赖范围广泛、有时相互冲突的国家法律几乎是不可能的。在全球范围内，随着技术的发展，我们需要围绕实时问题进行灵活的仲裁。

显然，仲裁机构将如何运作的具体细节必须进行谈判，也许这是科学机构的首要关注点之一。评审团需要多元化。它应该在一



个稳定的、政治中立的司法管辖区内拥有一个正式的地点或场所，并在立法上严格尊重仲裁裁决。它需要设立一个秘书处协助整理有关各方提交的证据和材料。它需要一种资金机制——最好有政府和工业界的支持，并实施合适的防火墙机制。至关重要的是，仲裁机构应当能够从广泛的利益相关者群体中收到意见——类似于“法庭之友摘要”在某些国家法院系统中所起的作用。非争议当事方的利益相关者可通过提供与所考虑问题相关的信息、专业知识和见解来协助仲裁过程。

仲裁庭将裁决公共、私营和民间部门之间的争端或分歧。例如围绕知识产权、公共服务提供的公共授权范围或公共机构使用人工智能的后果而产生的争端或分歧。经过我们的讨论，委员会决定，仲裁小组需要通过公共和私营部门参与者的合约或声明性“选择加入”来获得权力，以实现透明标准、可信评估、快速反应和全球一致裁决。至关重要的是，两个部门都同意受仲裁庭决定的约束，这意味着各国政府必须放弃主权豁免主张，并同意受到国内法的范围内规则和裁决的约束。同样，私人参与者必须同意遵守仲裁机构的决定，而不是通过向其他不太适合的机构或相关机构上诉寻求其他救济措施或替代结果。

仲裁法院的优势在于，它能作出积极有用的先例判决，通常能灵活地适用当下问题，并且可以比制定详细法律的程序更快地采取行动。在公共行政中使用人工智能的专门仲裁法庭将以高超的技术能力竭力解决分歧。

合作

人工智能如今的用途颇为广泛，但它在善治方面的应用缺乏政府、工业界、公民利益相关者和研究人员之间的包容、合作和参与机制。科学和仲裁机构之间更多的跨国合作可以促进对人工智能监管的信任，增加共识，并最终加强民众对人工智能如何在治理中实施和使用的支持。要求“更多的合作”是远远不够的。合作应该实现一定意义，与国际利益相关者独立沟通，以对话和共识为导向，同时采用包容性和针对性的方式。

欧洲或跨大西洋层面的联合政策倡议和目标、工业界合作、科学会议以及为以人工智能为重点的研究和宣传项目提供资金，为促进此类网络和知识交流提供了一个起点。但地缘政治差异和竞争压力阻碍了不同利益相关者群体和国界之间的公平合作。独立的仲裁制度和建立科学共识的机构有助于在国内和国际层面促进政府、工业界和民间社会之间的合作——或者至少是对话。

总之，如果资源充足，这两个组织将在未来几年的创新中提供全面的指导。这两个组织将作为一个集体发挥作用，一个致力于使用最新的科学方法生成证据，另一个致力于使用这些证据评估社会后果。这种组织方式在国际体系中有很好的先例。仲裁裁决可通过 1958 年《纽约仲裁公约》（《承认及执行外国仲裁裁决公约》）执行，该公约有 168 个缔约国。体育仲裁法院在快速决策和技术能力方面享有很高的国际声誉，其工作也赢得了广泛的信任。国际广告自律委员会和国际商会都是成熟的组织，它们制定标准，提供具有约束力的私人仲裁，并获得了联合国咨商地位。当然，需要更多的研究来建立法律框架，通过该框架，工业界、政府和民间社会可以通过私法进行合作。但关于这种合作的开展方式的榜样为我们带来了希望与鼓舞。

建议 1

在国际层面，政府、工业界和民间社会必须共同努力，建立和赋能 (a) 一个国际科学机构来推进公共服务人工智能应用的研究；以及 (b) 一个仲裁机构，以裁决公共服务人工智能系统中涉及的利益相关者之间可能出现的争议。

我们如何为人工智能建设公共服务能力，以实现善治？

人工智能、机器学习和其他先进的算法系统发展极为迅速，并且这一势头应该还会继续。人工智能几乎涵盖公共生活的所有领域，因此已成为监管的一个跨部门挑战。我们提议设立机构，制定原则和标准，并要求遵守这些原则和标准。但围绕人工智能在公共服务中的日常使用的问题往往琐碎得多，并且具有实际意义。我们需要如何培训公务员来有效监管公共服务中的人工智能?在本报告中，我们提出了一个广泛的监管概念，包括在原则或标准的授权范围内制定行政规则，例如拟议的科学机构制定的原则或标准。

首先，我们建议应用我们在第 3 节中阐述的原则，并建议对以下方面进行监管：人工智能系统的*设计*；公共机构获取或许可机器学习系统的*采购*流程；*实施*过程，包括公众咨询、宣传活动和信息获取；以及长期*当责*流程，用于整理公众对应用机器学习处理集体行动问题的长期后果的反馈。

许多国家/地区已经成功地在公共服务中采用人工智能——无论是否对设计、采购、实施和当责流程进行了监管。但我们的研究强调，公务员仍然普遍缺乏人工智能的采购、设计、评估和实施所需的能力。下一步是为公务员提供基本的专业知识，使他们能够就人工智能做出正确的决策。当然，公务员的教育和监管机构的技能提升是应对这一挑战的核心，但这些过程需要大量资源和时间。在此期间，提供最佳实践和简单决策的实用工具箱可能有助于解决一些最紧迫的挑战。政府必须建立一个中央的、方便访问的知识库，以便经过验证的战略和专业知识能够在不同的部门之间流动。

为此，至关重要的是监管机构具备制定所需咨询工具和合规机制的能力和权限。当然，人工智能设计、采购、实施和当责的部分监管工作可能会外包——并非所有工作都需要由政府来完成。公共机构应在前三个领域发挥带头作用，根据一些公共准则由第三方完成当责和认证工作。

最后，虽然确定应当监管的*内*容很有用，但值得注意的是，监管过程——以及相应而来的监管计划——必须有足够的资源和能力来支持专门研究公共部门人工智能问题的政府间工作人员团队。确保公共计划的管理人员训练有素至关重要，但同时应该在政府机构中普及技术知识，使政府能够更广泛地与其他工业界和民间社会利益相关者接触。

建议 2

在这两个新的国际组织的支持下，各国政府必须 (a) 建设其公共服务的能力，以深入参与公共服务人工智能系统的设计、采购、实施和当责；以及 (b) 为公共机构的工作人员提供工具箱，以完成其监督工作。

我们如何确保公共服务中的人工智能是值得信赖、可信赖的？

缺乏信任可能成为人工智能在公共服务中成功和及时实施的关键障碍：如果公众不信任人工智能，基于人工智能系统的政治决策必然会遭到强烈反对。当然，只有当技术系统确实值得信赖时，公众才应当信任人工智能。因此，牛津人工智能与善治委员会认为，加强公众支持和人工智能评估对于实施人工智能以实现善治至关重要。我们建议政府和公共机构解决民众对人工智能使用的担忧，并采取以教育、扫盲和认证为目标的措施来解决这些问题。

首先，为了加强公众对公共服务人工智能的信任，我们建议各国政府开展公共宣传和扫盲活动，强调人工智能在公众已经熟悉的领域在公共利益问题方面的善用。宣传活动还应考虑到围绕系统缺陷和政府对人工智能系统的滥用的现实关切点。我们建议各国政府为及时、跨媒体的宣传活动和营销投入大量的宣传预算，以覆盖不同的人口群体。随着关于人工智能的负面消息和误导性信息变得普遍，阐明人工智能在早期诊断、交通管理和气候技术等领域的使用可以提高民众对人工智能技术的认可。

第二，从长远来看，为了为这一领域的进一步工作指明方向，各国政府应努力更好地披露何时以及如何使用人工智能或计划使用人工智能。披露人工智能在公共服务中的使用方面和使用方式将有助于提高透明度并最终带来对人工智能的信任。政府网页和中央数据库上的披露说明，以及可公开访问的人工智能供应商和政府开发的人工智能系统名单，可以作为全面披露制度的首要工作。还可以由我们推荐的科学机构构建一个综合的国际数据库，这也是很有利的举措。

第三，我们提出了在公共服务中使用人工智能的全球认证体系。此类人工智能认证应考虑安全和质量指标，并公开政府使用被认为尚不完善的人工智能系统的记录。关于善意和值得信赖的人工智能的可操作化的学术辩论得到了很好的发展，研究人员提出了规范的伦理框架和可衡量的指标来评估人工智能系统的透明度、可解释性和当责制。因此，质量、风险和影响评估、设计过程中的严格测试、培训数据评估和系统维护在工业界得到了更广泛的采用，但尚未形成全面的最佳实践或标准。我们完全不必担心认证会阻碍创新。世界各地都有针对金融交易软件、电子赌博机和数据处理的严格审计和认证实践。设计全面的认证系统和政策工具箱可能是拟议的科学机构的首要任务之一，尽管最终最适合进行认证的可能是国际标准机构或工业机构。

建议 3

必须通过关于日常应用和即将出现的实际用例、其影响和风险的公共教育活动来加强人们对人工智能使用的信任，方法是 (a) 要求政府公开披露人工智能技术在公共服务中的使用情况；以及 (b) 引入一个多部门机构，由其提供一个基本的认证系统，不断验证应用程序，以随着时间的推移建立信任。

5 结语：立即行动

确定广泛的原则对于制定理解公共问题的框架至关重要。有助于我们评估各种挑战，并确定具体的行动建议。

介绍了本报告中提出的三项建议之后，我们现在来谈一谈下一步工作。了解了我们对将人工智能投入公共服务所面临的挑战，并就如何通过人工智能系统支持善治提出了一系列建议，下一步应该怎么做？

I. 组织可行性研究

开展可行性研究，以估算组织所需仲裁和科学机构的成本。这两个机构需要在有利于其运作的司法管辖区和政治环境中有固定场所，这一点毋庸置疑。谁将领导这些组织，以及需要什么样的具体组织能力来完成他们的重要使命？要采取的第一步是准备一份可行性研究报告，涉及组织管理和运作场景，以建立我们将人工智能投入公共服务的全球能力。

II. 与现有的国家和多边机构进行沟通协商

许多高效处理各类事项的多边机构为关于人工智能在公共服务领域应用的全球对话做出了重要贡献。然而，这些机构中没有任何一个拥有广泛的权限，也没有一个将这项工作视为其核心、专门的使命。此外，需要完成的工作的几个方面不属于现有多边机构的任务范围。联合国教科文组织的专家已经看到了人工智能应用的许多重要趋势，联合国开发计划署以及 G7（七国集团）和 G20（二十国集团）已经开始讨论人工智能在治理中的作用。联合国人工智能机构间工作组会定期举行会议，会议由联合国教科文组织和国际电信联盟领导。

我们认为，全球政策制定者和决策者需要为实现更有意义的合作铺平道路。G7 的全球人工智能伙伴关系 (GPAI) 是朝着这个方向做出的一项值得称赞的努力，它将代表多方利益相关者的来自科学、技术、民间社会和政策领域的专家聚集在一起。但 GPAI 是一个相当封闭的团体，缺乏人工智能领域一些最重要的创新者以及南方国家的代表。值得注意的是，对于与使用人工智能相关的侵犯人权的更广泛的讨论和担忧，作为人工智能领域全球领导者的中国一直存在感不强。根据我们在 OxCAIGG 的研究和专家讨论，我们强调了更多国际合作和沟通交流的重要性。

III. 国际里程碑事件中的包容性对话

对于即将发生的许多重要事件，可以就提高我们确保人工智能用于善治的能力的现实前景进行包容性对话。例如，2023 年 9 月，联合国将举办未来峰会，国家元首作为全球政策制定者与会。这一活动可以作为在中立地区进行对话的关键，也可以作为更持续交流的起点，包括在有边际共识的利益相关者群体之间进行交流。这种规模的全球活动很有可能促成更多的长期系统合作。

目标摘要

建议

1

在国际层面，政府、工业界和民间社会必须共同努力，建立和 赋能

(a) 一个国际科学机构来推进公共服务人工智能应用的研究；以及

(b) 一个仲裁机构，以裁决公共服务人工智能系统中涉及的利益相关者之间可能出现的争议。

2

在这两个新的国际组织的支持下，各国政府必须

(a) 建设其公共服务的能力，以深入参与公共服务人工智能系统的设计、采购、实施和当责；以及

(b) 为公共机构的工作人员提供工具箱，以完成其监督工作。

3

必须通过关于日常应用和即将出现的实际用例、其影响和风险的公共教育活动来加强人们对人工智能使用的信任，方法是 (a) 要求政府公开披露人工智能技术在公共服务中的使用情况；以及

(b) 引入一个多部门机构，由其提供一个基本的认证系统，不断验证应用程序，以随着时间的推移建立信任。

立即行动的下一步工作

I

组织可行性研究

II

与现有的国家和多边机构进行沟通协商

III

国际里程碑事件中的包容性对话

6.OxCAIGG 开展研究的参考文献

[1] Neudert, L.-M., Knuutila, A. & Howard, P. N. *全球对人工智能、机器学习和自动化决策的态度*。(研究报告，2020 年 10 月，牛津人工智能与善治委员会，2020 年)。

[2] Ramizo, G., Jr. *政府人工智能项目的实践经验*。(研究报告，2021 年 1 月，牛津人工智能与善治委员会，2021 年)。

[3] Vogl, T. *地方政府与人工智能*。(研究报告，2021 年 2 月，牛津人工智能与善治委员会，2021 年)。

[4] Nyabola, N. *老问题，新技术：高度分裂和社会化分层的社会中的人工智能、人权和善治。肯尼亚的案例*。(研究报告，2021 年 3 月，牛津人工智能与善治委员会，2021 年)。

[5] Au, Y. *监控即服务：欧洲人工智能辅助监控系统的大规模市场*。(研究报告，2021 年 4 月，牛津人工智能与善治委员会，2021 年)。

[6] McFadden, M., Jones, K., Taylor, E. & Osborn, G. *人工智能的协调：标准在欧盟人工智能法规中的作用*。(研究报告，2021 年 5 月，牛津人工智能与善治委员会，2021 年)。

[7] Neudert, L.-M. & Howard, P. N. *人工智能与善治整合的四项原则*。(研究报告，2020 年 1 月，牛津人工智能与善治委员会，2020 年)。

委员会简介

将人工智能用于实现善治和公共服务的挑战是世界各国迫切关注的问题。牛津人工智能与善治委员会于 2020 年 7 月成立，其目标是制定原则和切实可行的政策建议，确保民主国家借助人工智能实现善治。

最近，受新冠疫情的影响，人工智能解决方案激增。虽然这些新技术是为了实现公共利益，但这些新技术在评估这些产品的适用性和合法性方面带来了挑战。人工智能系统的实施速度是前所未有的，这表明政府需要围绕这些类型的人工智能产品及其采购和实施来制定政策。

OxCAIGG 研究了全球民主国家在利用人工智能实现善治方面面临的采购和实施挑战，确定了评估和管理风险和收效的最佳实践，并提出了一些战略建议，旨在充分利用技术能力，同时减轻支持人工智能的公共政策的潜在危害。

根据来自各地区和各专业领域的专家的意见，包括来自政府、工业界、技术协会和民间社会的利益相关者，OxCAIGG 为使用人工智能实现善治提出了适用的和相关的建议。

OxCAIGG 的委员运用他们的经验和洞察力，经过精心考虑，为政策制定者提供指导并为其指明方向，以确保在不久的将来适应和采用人工智能相关工具以实现善治。

委员简介

Yuichiro Anzai（安西佑一郎）博士（OxCAIGG 委员）

日本学术振兴会高级顾问和庆应义塾大学教务处执行顾问。

作为人工智能战略委员会的主席，安西佑一郎负责为日本政府制定战略政策提供咨询意见。他因在整合认知科学和信息科学方面的开创性工作而荣获日本政府授予的“文化功勋人物”称号。

Rumman Chowdhury 博士（OxCAIGG 委员）

Twitter META（机器学习道德规范、透明和责任）部门主任。

Rumman Chowdhury 痴迷于人工智能与人之间的交集。她是应用算法伦理领域的先驱，为符合伦理、可解释和透明的人工智能创造了尖端的社会/技术解决方案。

Tom Fletcher 先生（CMG，OxCAIGG 委员）

牛津大学赫特福德学院院长，Foundation for Opportunity 基金创始人。

Tom Fletcher 曾任三位英国首相的外交政策顾问；英国驻黎巴嫩大使；纽约大学客座教授；全球商业教育联盟顾问；并担任英国创意产业联盟国际委员会主席。他还是《The Naked Diplomat》一书的作者。

女爵士 Wendy Hall 教授（大英帝国女爵士，皇家学会会员，皇家工程学院院士，OxCAIGG 委员）

南安普敦大学计算机科学系钦定教授，副院长（国际事务），工程与科学学院执行主任。

Wendy Hall 是 Ada Lovelace 研究所的主席，也是 BT 技术咨询委员会的成员。她在 2009 年的“英国新年荣誉榜”上荣获大英帝国爵级司令勋章（女），并且是皇家学会的会员。

Philip Howard 教授（OxCAIGG 委员）

牛津大学贝利奥尔学院网络研究首席教授。

Philip Howard 研究数字媒体对世界各地政治生活的影响，并且经常担任全球媒体和政治事务的评论员。他是牛津大学民主与技术项目的负责人，该项目研究算法和自动化在日常生活中的使用。

Julian King 爵士（OxCAIGG 委员）

前欧盟委员和英国外交官。

Julian King 曾担任英国驻爱尔兰和法国大使以及北爱尔兰办事处主任，也是最后一位在欧盟委员会担任安全联盟委员的英国官员。

Safiya Noble 博士（OxCAIGG 委员）

加州大学洛杉矶分校 (UCLA) 非裔美国人研究系和信息研究系副教授，加州大学洛杉矶分校关键互联网查询中心 (C2i2) 联合主任。

Safiya Noble 的工作涉及社会学，同时又是跨学科的，标记了数字媒体对种族、性别、文化和技术问题的影响和交叉方式。她是网络民权计划的董事会成员，为那些容易受到网络骚扰的人提供服务。

Howard Rosen 先生（大英帝国司令勋章，OxCAIGG 委员）

Howard Rosen Solicitors 律师事务所负责人，牛津大学瑞士之友会主席，Rail Working Group 主席，英联邦犹太人委员会托管人。

Howard Rosen 专注于国际商业、金融和租赁法律和信托等领域。他还是瑞士楚格信托公司 Rosetrust AG 和瑞士尼翁的 Aviation Advocacy Sàrl 的创始人兼董事总经理。

申卫星教授（OxCAIGG 委员）

清华大学法学院院长、清华大学智能法治研究院院长、清华大学计算法学硕士项目负责人。

申卫星是中国人工智能产业发展联盟理事，中国法学会常务理事，中国法学会网络与信息法学研究会副会长。资助并发表了两篇关于人工智能和善治的当代问题的社论。

Joanna Shields 女男爵（大英帝国官佐勋章，OxCAIGG 委员）

BenevolentAI 首席执行官

Joanna Shields 是科技行业的资深人士，曾成功创办了几家世界知名公司。她对造福人类的科技非常热衷。她有着 30 多年的职业经验，专注于利用技术的力量推动变革，提高互联，强调人与社会的力量。

鸣谢

感谢 Flora Seddon 在过去 18 个月中对 OxCAIGG 活动的协调与支持。感谢 Bruno Selun 和 Kumquat 的团队对我们的数字咨询和战略对话的支持。感谢 Hubert Au、Rutendo Chabikwa、Tim Curnow 博士、Mona Elswah、John Gilbert、Mark Healy、Lucy Hennings 博士、Mark Malbas、Nahema Marchal 博士、Sara Spinks 和 Niamh Walsh 的贡献。感谢为本报告和 OxCAIGG 做出贡献的专家、研究人员和公共服务专家。

感谢 Adessium Foundation、Civitates、Adessium Foundation、Luminate 和 Open Society Foundations 等基金会对 OxCAIGG 的支持。本报告中的任何意见、结果、结论或建议均属委员会观点，不一定代表牛津大学、我们的资助者或个别委员的观点。牛津大学的中央大学研究伦理委员会负责编写牛津大学研究的伦理监督报告，每项研究成果的审批编号请参见相关的研究报告。

oxcaigg.oii.ox.ac.uk

牛津人工智能与善治委员会。
公共服务中的人工智能：从原则到实践。
研究报告，2021 年 6 月，英国牛津：
牛津人工智能与善治委员会。19 页。

检索自：https://oxcaigg.oii.ox.ac.uk