智源学者成果展示——人工智能伦理与安全

作者　曾毅 Yi Zeng
中科院自动化所中英人工智能伦理与治理研究中心
北京智源人工智能研究院人工智能伦理与安全中心
China-UK Research Centre for AI Ethics and Governance,
Institute of Automation, Chinese Academy of Sciences
Research Center for Artificial Intelligence Ethics and Safety,
Beijing Academy of Artificial Intelligence

2020 年 6 月

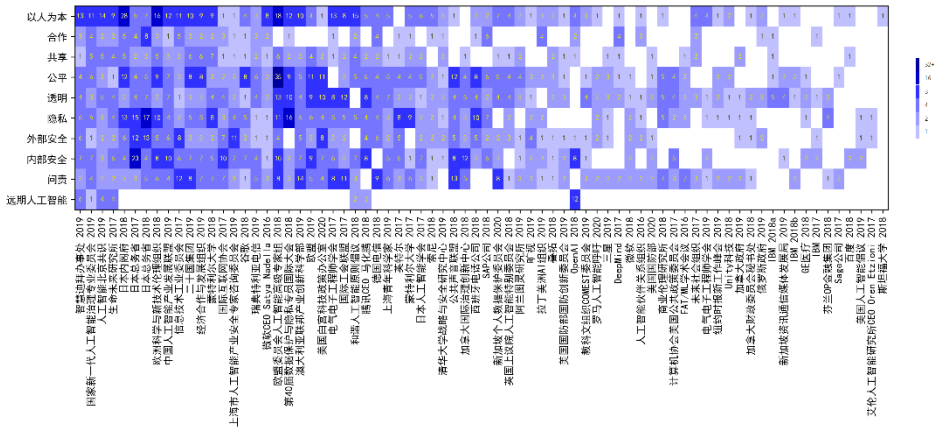
LAIP – 链接人工智能准则平台

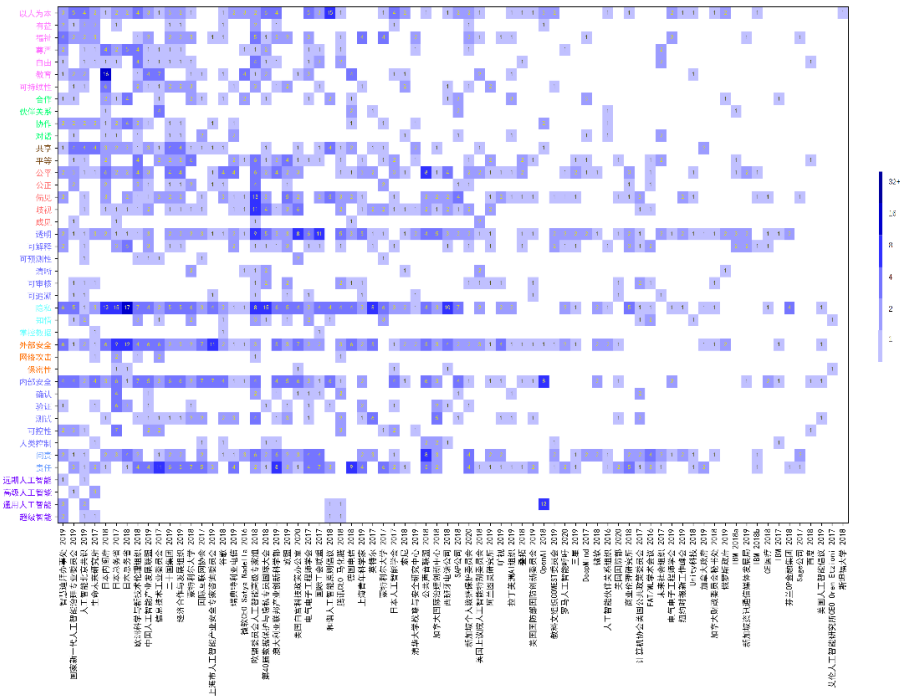
每个人工智能准则背后，都有不同的考虑，故其中任何一个都不可能完美适用于任何场景。 链接人工智能准则平台 (Linking Artificial Intelligence Principles，LAIP) 的任务是连接、整合与分析世界各地的人工智能准则。这些人工智能准则来自不同的研究机构，非营利机构，非政府组织，公司等等。平台致力于研究人工智能准则之间的相通之处，不同之处与互补之处。

当前链接人工智能准则平台可以让我们在关键词，主题与篇章的层次列举和比较不同的人工智能准则。

【 网站链接 】

http://www.linking-ai-principles.org/

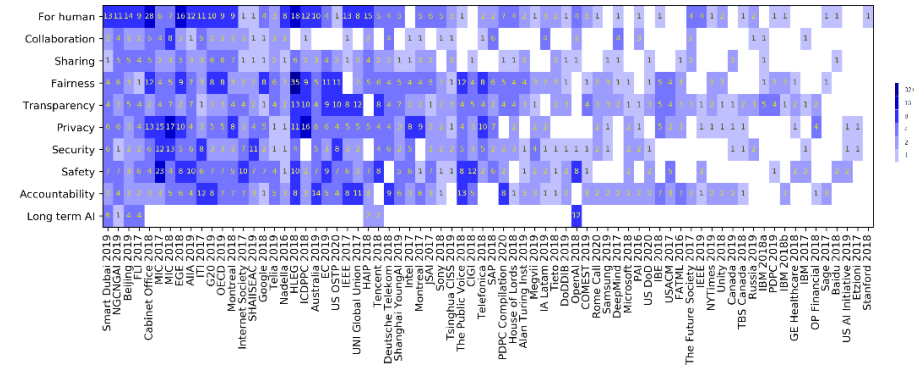


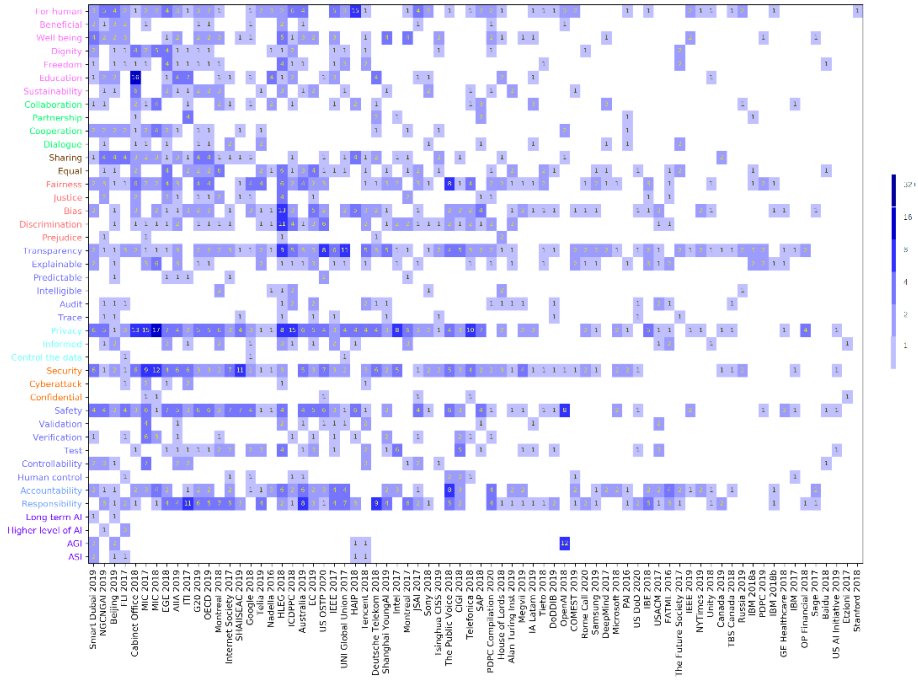


LAIP – Linking Artificial Intelligence Principles

Various Artificial Intelligence Principles are designed with different considerations, and none of them can be perfect and complete for every scenario. Linking Artificial Intelligence Principles (LAIP) is an initiative and platform for integrating, synthesizing, analyzing, and promoting global Artificial Intelligence Principles and their social and technical practices World Wide, from different research institutes, non-profit organizations, non-governmental organizations, companies, etc. The efforts aim at understanding in which degree do these different AI Principles proposals share common values, differ and complete each other.

The current LAIP engine enables us to list and compare between different AI principle proposals at keywords, topic and paragraph levels.





人工智能治理沙盒

人工智能治理沙盒
Artificial Intelligence Governance Sandbox

依据世界众多AI原则测评您的AI项目

在线测评后将自动生成报告，内容包括基于全球AI治理原则及其详细解释，您的项目中应当注意和改善的部分。

开始测评

依次点击每一个主题，回答所有问题来完成测评

+ 合法合规

+ 环境与社会有益

2.1. 该系统的设计与应用中，是否充分考虑了其对环境和社会可持续发展的影响？

联合国所有成员国于2015年正式通过了17个可持续发展目标，这些目标涵盖了消除贫困和饥饿、改善健康和教育、减少不平等、促进经济增长、应对气候变化、保护森林与海洋、推进法治与人权保护、加强各方协作等。这些可持续发展目标为全球2015-2030年的社会、经济和环境发展指明了方向。

- ○ 不涉及这些方面
- ◉ 尚未考虑
- ○ 已考虑

2.2. 该系统的部署是否有可能会加剧相应行业的数据或平台垄断？或者有助于避免上述数据或平台垄断的出现？

- ○ 没有明显影响
- ◉ 可能会加剧相应行业的数据或平台垄断
- ○ 有助于避免相应行业的数据或平台垄断

2.3. 该系统的广泛部署是否可能会造成特定群体的技术性失业？如果是，是否可以通过诸如替代就业、培训教育等方式使上述影响可控？

- ○ 不会对任何群体的就业产生明显影响

问题 7.4. 该系统在开发、测试和部署时，在收集、使用用户个人信息的过程中是否遵守了与用户的其他约定?

您的选择是: ' 否'

解析:

根据**《中华人民共和国网络安全法》**，第四十一条 网络运营者收集、使用个人信息,应当遵循合法、正当、必要的原则,公开收集、使用规则,明示收集、使用信息的目的、方式和范围,并经被收集者同意。 网络运营者不得收集与其提供的服务无关的个人信息,不得违反法律、行政法规的规定和双方的约定收集、使用个人信息,并应当依照法律、行政法规的规定和与用户的约定,处理其保存的个人信息

根据**《新一代人工智能治理原则》** 四、尊重隐私。人工智能发展应尊重和保护个人隐私，充分保障个人的知情权和选择权。在个人信息的收集、存储、处理、使用等各环节应设置边界，建立规范。完善个人数据授权撤销机制，反对任何窃取、篡改、泄露和其他非法收集利用个人信息的行为

根据**《人工智能北京共识》** 服务于人：人工智能的研发应服务于人类，符合人类价值观，符合人类的整体利益; 应充分尊重人类的隐私、尊严、自由、自主、权利; 人工智能不应被用来针对、利用或伤害人类

问题 7.5. 该系统在开发、测试和部署时，对于所收集的用户个人信息是否进行了充足的安全保障（包括制度上及技术上）来防范数据被窃取、篡改、泄露或其他非法行为利用？相关措施的效果如何?

您的选择是: ' 采取了措施，但未能实现充分安全保障'

解析:

根据**《中华人民共和国网络安全法》**，第四十二条 网络运营者不得泄露、篡改、毁损其收集的个人信息;未经被收集者同意,不得向他人提供个人信息。但是,经过处理无法识别特定个人且不能复原的除外。 网络运营者应当采取技术措施和其他必要措施,确保其收集的个人信息安全,防止信息泄露、毁损、丢失。在发生或者可能发生个人信息泄露、毁损、丢失的情况时,应当立即采取补救措施,按照规定及时告知用户并向有关主管部门报告，第四十四条 任何个人和组织不得窃取或者以其他非法方式获取个人信息,不得非法出售或者非法向他人提供个人信息

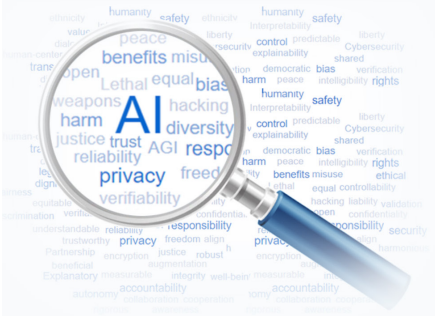
根据**《新一代人工智能治理原则》** 四、尊重隐私。人工智能发展应尊重和保护个人隐私，充分保障个人的知情权和选择权。在个人信息的收集、存储、处理、使用等各环节应设置边界，建立规范。完善个人数据授权撤销机制，反对任何窃取、篡改、泄露和其他非法收集利用个人信息的行为

根据**《人工智能北京共识》** 服务于人：人工智能的研发应服务于人类，符合人类价值观，符合人类的整体利益; 应充分尊重人类的隐私、尊严、自由、自主、权利; 人工智能不应被用来针对、利用或伤害人类。 控制风险：人工智能及其产品的研发者应

【网站链接】

http://ai-governance-sandbox.org/

Artificial Intelligence Governance Sandbox



Artificial Intelligence Governance Sandbox

Evaluate your AI project based on comprehensive AI principles and norms World Wide.

An automated report suggesting where should be noticed and improved in your project based on global AI governance principles and detailed explanation will be generated immediately after the online evaluation.

Start Evaluation

To start your evaluation
Please click each topic and answer all of the evaluation questions.

+ Legal compliance

+ Environmental and social well-being

2.1. Has the impact on environmental and social sustainability been fully considered in the design and application of the system?

A set of 17 Sustainable Development Goals (SDGs) has been adopted by all United Nations Member States in 2015, which covers issues such as ending poverty, improving healthcare and education, spurring economic growth, reducing inequality, tackling climate change, working to preserve oceans and forests, defending justice and human rights, and strengthening partnerships. These SDGs set the direction for the global social, economic and environmental development in 2015-2030.

- ○ N/A (Not related to these aspects)
- ◉ No such considerations yet
- ○ Has been considered

2.2. Is the deployment of the system likely to exacerbate current data/platform monopolies in the relevant industry? Or will it help avoid such data/platform monopolies?

- ○ N/A (No significant impact)
- ◉ May exacerbate data/platform monopolies
- ○ Helps to avoid data/platform monopolies

Question 2.1. Has the impact on environmental and social sustainability been fully considered in the design and application of the system?

Your choice: ' No such considerations yet'

Explanation:

According to the Governance Principles for the New Generation Artificial Intelligence, In order to promote the healthy development of the new generation of AI, better balance between development and governance, ensure the safety, reliability and controllability of AI, support the economic, social, and environmental pillars of the UN sustainable development goals, and to jointly build a human community with a shared future, all stakeholders concerned with AI development should observe the following principles: ...

According to the Beijing AI Principles, Do Good: AI should be designed and developed to promote the progress of society and human civilization, to promote the sustainable development of nature and society, to benefit all mankind and the environment, and to enhance the well-being of society and ecology.

Show more discussions based on other AI principles

According to the OECD Principles on Artificial Intelligence, 1.1. Inclusive growth, sustainable development and well-being: Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.

According to the Key requirements for trustworthy AI, VI. Societal and environmental well-being: For AI to be trustworthy, its impact on the environment and other sentient beings should be taken into account. Ideally, all humans, including future generations, should benefit from biodiversity and a habitable environment. Sustainability and ecological responsibility of AI systems should hence be encouraged. The same applies to AI solutions addressing areas of global concern, such as for instance the UN Sustainable Development Goals. Furthermore, the impact of AI systems should be considered not only from an individual perspective, but also from the perspective of society as a whole. The use of AI systems should be given careful consideration particularly in situations relating to the democratic process, including opinion-formation, political decision-making or electoral contexts. Moreover, AI's social impact should be considered. While AI systems can be used to enhance social skills, they can equally contribute to their deterioration.

According to the Social Principles of Human-centric AI (Draft), Fundamental Philosophy: ... We consider it to be

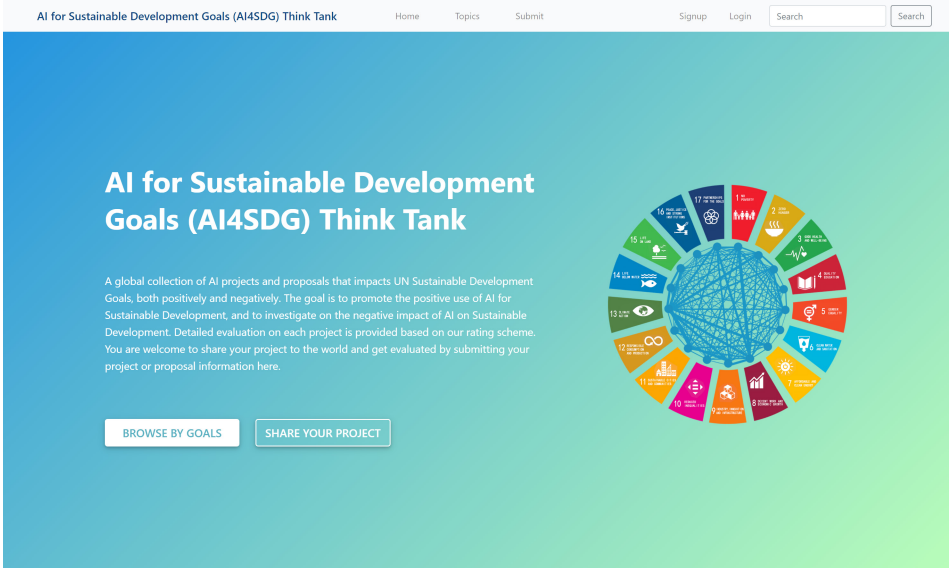
AI for Sustainable Development Goals (AI4SDG) Think Tank

AI for Sustainable Development Goals (AI4SDG) Think Tank

A global collection of AI projects and proposals that impacts UN Sustainable Development Goals, both positively and negatively. The goal is to promote the positive use of AI for Sustainable Development, and to investigate on the negative impact of AI on Sustainable Development. Detailed evaluation on each project is provided based on our rating scheme. You are welcome to share your project to the world and get evaluated by submitting your project or proposal information here.

【网站链接】

http://ai-for-sdgs.academy/



Browse by UN Sustainable Development Goals

Recent Projects

AI's killer (whale) app
United States
Goal 14 Life Below Water

NatureServe Habitat Suitability Modeling
USA
Goal 15 Life on Land

AI-BASED EARLY PEST WARNING SYSTEM
India
Goal 2 Zero Hunger

Using AI to find where the wild things are
Goal 15 Life on Land

OOICloud
Goal 14 Life Below Water

Aerobotics

Brief Project Information

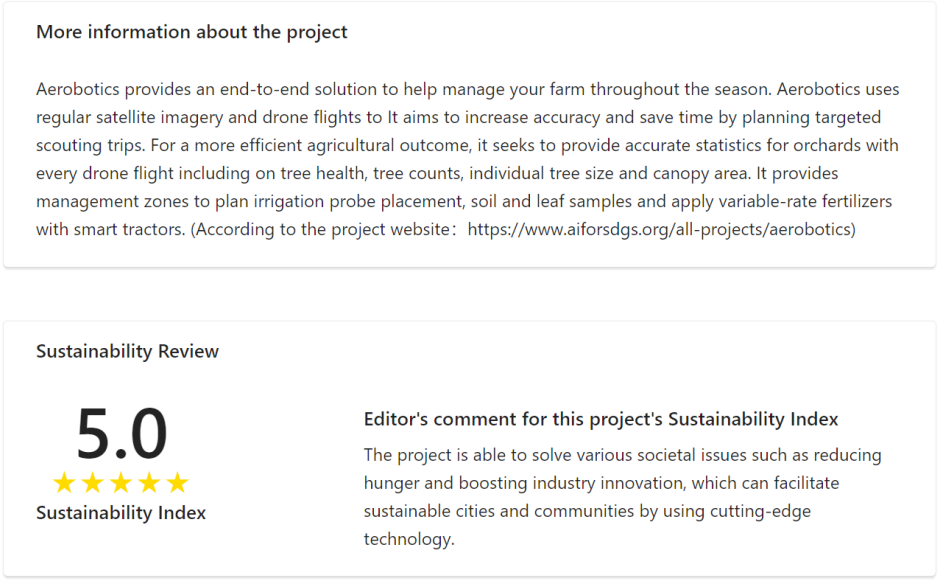
Aerobotics uses drone imagery and artificial intelligence to monitor crops and warn about potential risks, early pest, and disease detection.

SDG

Project Link

https://aerobotics.co/

South Africa



More information about the project

Aerobotics provides an end-to-end solution to help manage your farm throughout the season. Aerobotics uses regular satellite imagery and drone flights to It aims to increase accuracy and save time by planning targeted scouting trips. For a more efficient agricultural outcome, it seeks to provide accurate statistics for orchards with every drone flight including on tree health, tree counts, individual tree size and canopy area. It provides management zones to plan irrigation probe placement, soil and leaf samples and apply variable-rate fertilizers with smart tractors. (According to the project website：https://www.aiforsdgs.org/all-projects/aerobotics)

Sustainability Review

5.0

★★★★★

Sustainability Index

Editor's comment for this project's Sustainability Index

The project is able to solve various societal issues such as reducing hunger and boosting industry innovation, which can facilitate sustainable cities and communities by using cutting-edge technology.

人工智能伦理、治理与可持续发展译丛

【网站链接】

https://www.baai.ac.cn/research/translation-series-on-ai-ethics

译丛引言

人工智能伦理与治理关乎全球人工智能发展与创新的方向与未来。将人工智能作为使能技术推动人类、社会、生态及全球可持续发展是人类进行人工智能技术创新的共同愿景。在这个过程中，来自各个国家、政府间组织、国际组织的人工智能伦理与治理工作通过学术机构、产业、政府等以各种方式积极推动相关原则、政策、标准、法律的制定、技术与社会落地。

虽然来自各个国家、组织的努力是在不同文化背景下建立的，但是文化的差异恰恰提供给我们思考问题的不同视角，和相互学习与借鉴的机会。如《论语》中有言“君子和而不同”。建立跨文化互信是全球和谐发展的基石。人工智能伦理、治理与可持续

发展将是全球科技、社会领域的持续性重要议题。

为此，北京智源人工智能研究院人工智能伦理与安全研究中心携手中国科学院自动化研究所中英人工智能伦理与治理研究中心等单位共同发起《人工智能伦理、治理与可持续发展 译丛》，将人工智能伦理与治理、可持续发展领域的重要文献进行遴选，组织翻译，并介绍给全球读者。期待从跨文化、跨语言的交流中各自有所裨益，促进伴随技术发展的文化交流，推动全球人工智能与人类未来的和谐发展。

曾毅

北京智源人工智能研究院人工智能伦理与安全研究中心 主任

中国科学院自动化研究所中英人工智能伦理与治理研究中心 主任

Seán Ó hÉigeartaigh

剑桥大学生存风险研究中心 联合主任

剑桥大学未来智能研究中心研究中心 研究主任

【下面是部分文档的封面展示】

指导数字近距离追踪技术用于2019冠状病毒病（COVID-19）接触者追踪的伦理考虑

临时指导文件
2020年5月28日

世界卫生组织

背景

本临时指导文件旨在为正在考虑是否需要为2019冠状病毒病（COVID-19）接触者追踪开发或实施数字近距离追踪技术的公共卫生方案和政府提供信息。本文件涵盖了伦理原则、符合这些原则的技术考虑和要求；以及如何公平合理地使用此类技术。

接触者追踪是识别、评估和管理接触过某种疾病的人，以防止疾病继续传播的过程。系统地实施接触者追踪可以切断传染病的传播链，因而是控制传染病暴发的必要的公共卫生工具。为了有效地进行接触者追踪，各国必须有足够的能力，包括人力资源，以及时检测疑似病例。[1]数字技术可以在成员国实施的接触者追踪方案中发挥作用。

根据《国际卫生条例》，成员国有义务开发公共卫生监测系统，[2]为本国应对COVID-19采集关键数据，同时确保此类系统是透明的，能够响应社区的关切，并且不会造成如侵犯隐私权等不必要的负担。[3]未能有效实施监测系统可能会妨碍公共卫生和临床应对措施的有效进行。[4]数字技术被用于公共卫生监测，以支持快速报告、数据管理和分析；尤其是与机器学习和人工智能相结合时，可以构成强大的工具，为公共卫生机构提供有价值的信息，以做出适当的决策。[5]

近几个月来，近距离追踪作为一种用于监测的数字技术在多个COVID-19疫情国受到关注。近距离追踪技术通过测量信号强度来确定两个设备（如智能手机）是否相互在近距离内，该距离足以致使设备用户将病毒从感染者传播给未感染者。如果某用户被感染，被识别位于该用户周围近距离的其他用户能够收到通知，从而采取适当措施降低自己和他人的健康风险。[6]近距离追踪与"接触者追踪"经常被混为一谈，但是，接触者追踪是一门广泛的公共卫生学科，而近距离追踪则是辅助接触者追踪的一项新技术。

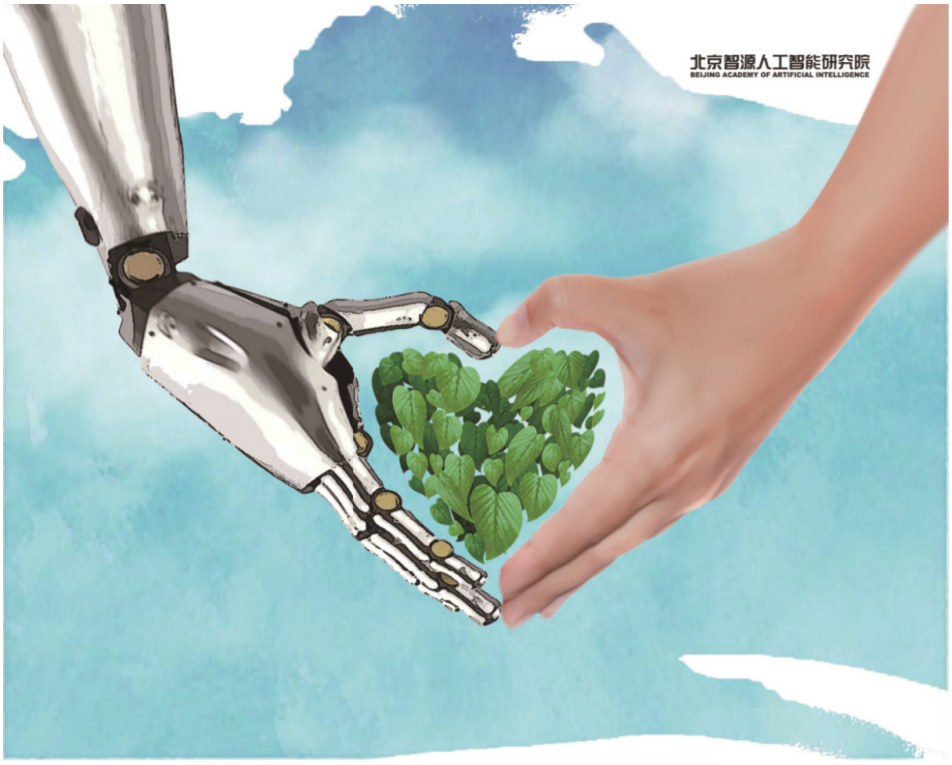
然而，数字近距离追踪技术有其局限性。该技术无法采集可能会导致用户感染COVID-19的所有场景，且不能替代传统的人对人的公共卫生追踪、检测或外展，后者通常通过电话或面对面进行。只有将数字近距离追踪应用程序完全纳入现有的公共卫生系统和全国性大流行病应对措施中，才能有效提供数据以帮助应对COVID-19疫情。这样的系统需要包括医疗卫生服务人员、检测服务和手动追踪接触者的基础设施。[7]

鉴于这些局限性，卫生部门可以在某人接触COVID-19检测呈阳性的人员的风险增高时利用数字近距离追踪工具向其发出通知。向可能曾与COVID-19检测呈阳性的人员密切接触的人发布通知也许会促使其甚至早在出现任何症状前就寻求受检（如有检测条件）或采取自我隔离和保持身体距离等预防措施来抑制潜在传播。[8]早期的公共卫生应对措施可能会对COVID-19疫情是否能够得到控制抑或面临新一轮暴发产生重大影响。此外，由数字近距离追踪技术生成的数据可能有助于研究人员为今后可能发生的COVID-19暴发做好准备，也有助于为未来的流行病和大流行病做好总体准备。

然而，使用此类数据也可能在COVID-19大流行病期间以及之后威胁基本人权和自由。监测可能会很快越过疾病监测和人群监控之间的模糊界限。因此，需要法律、政策和监督机制来严格限制数字近距离追踪技术的应用以及任何使用由该技术生成的数据的研究。

一些私营公司通过其产品、服务或平台采集的数据量与政府收集的数据量相当。这些公司可能会开发甚至正在与政府共享其数字近距离追踪应用程序，在某些情况下，这些公司被授责收集并分析由此获得的数据。此外，人们日益担心私营公司可能会将其商业产品、服务和构架永久融入公共卫生基础设施中。

注：世界卫生组织目前并未提供官方翻译版本，本中文版本由北京智源人工智能研究院人工智能伦理与安全研究中心与中国科学院自动化研究所中英人工智能伦理与治理研究中心组织翻译，供相关方参考。中文翻译如与英文原文意义有不一致，请以英文原文为准。中文译文问题可联系 yi.zeng@ia.ac.cn
英文原文地址：https://www.who.int/publications-detail/WHO-2019-nCoV-Ethics_Contact_tracing_apps-2020.1



Facial Recognition and Public Heath

— The First Report in Survey Series on
Artificial Intelligence and Healthy Society

May 17th , 2020

Published by:

Research Center for AI Ethics and Safety, Beijing Academy of Artificial Intelligence

China-UK Research Centre for AI Ethics and Governance, Institute of Automation, Chinese Academy of Sciences

Beijing Academy of Artificial Intelligence

微信关注
北京智源人工智能研究院