08 人工智能伦理、治理与可持续发展

联合国副秘书长 Fabrizio Hochschild：联合国数字合作路线图

整理：智源社区 沈磊贤

6 月 24 日下午，在第二届北京智源大会“人工智能伦理、治理与可持续发展”论坛上，联合国副秘书长 Fabrizio Hochschild 连线出席，并做了《Digital Cooperation in the UN: The SG's Roadmap and implementation》的演讲。

Fabrizio Hochschild 介绍，《联合国数字合作路线图》的首要目标是“连接、尊重和保护数字时代的人们”，在全球新冠疫情大流行的当下，实现可持续和包容性复苏，推动全球数字合作，以应对在世界范围内出现的技术被严重滥用等问题，其主要内容包括推动数字通用连接，促进数字技术成为公共产品，保证数字技术惠及所有人，支持数字能力建设，建立数字信任和安全等。

Fabrizio Hochschild 介绍，这个路线图是由联合国秘书长古特雷斯 6 月 11 日在“数字世界现状和数字合作路线图实施”视频高级别活动上正式公布的，涉及诸如确保数字通用连接，通过网络维护人权，改善互联网合作架构等时下的重要议题，旨在促进全球数字合作。Fabrizio 认为，当下全球的数字合作正处于历史较低水平，但面对疫情，数字合作应该享有最高优先级，各国应该通力合作，推动数字合作达到历史最高水平。Fabrizio 希望更多的中国专家能够与联合国一起努力，促进数字合作，实现可持续发展的目标，保证数字技术能够惠及所有人，推动全球各国共同进步。

《联合国数字合作路线图》呼吁各国要在如下八个重要的数字领域采取合作，包括:

- 实现全球的互联互通，到 2030 年，每个人都应能够以安全且负担得起的方式访问互联网;
- 促进数字公共物品以创造更公平的世界;
- 确保所有人（包括最弱势群体）享有数字包容;
- 加强数字能力建设;
- 确保数字时代的人权保护;
- 支持人工智能方面的全球合作;
- 促进数字环境中的信任与安全;
- 建立更高效的数字合作架构。

这份数字合作路线图，也是由阿里巴巴马云共同主持的数字合作高级小组在去年发布报告（编者注:《数字相互依存的时代——联合国数字合作高级别小组报告》，2019 年 6 月发布）的后续行动，在多方利益相关者讨论的基础上制定。能够在数字合作项目的一开始就从中国的智慧和经验中受益，Fabrizio 表示很幸运。

而联合国也将在后续工作中寻求更多的中国专业知识，Fabrizio 希望能够接触更多尊重数字人权、保护数字时代弱势群体的中国专家。同时 Fabrizio 也提到了眼下全球面临的共同威胁与挑战：Covid-19 新冠疫情，其后果是毁灭性的，影响范围波及世界各地，加剧了全球紧张局势。

尽管现实世界存在着复杂的地缘政策，但最近几个月的发展趋势，需要我们强调采取技术合作的迫切需要。Fabrizio 提到，在全球最不发达国家中，有 46%的人口无法享受互联网访问带来的好处。如何使这些人口能够从人工智能等先进技术中受益？ Fabrizio 邀请与会嘉宾与他一起思考这一困境以及解决这一困境的方法。

Fabrizio 认为，造成全球政策分裂的最大风险之一是技术分裂，技术分裂将使全世界人民更难平等地获得先进的技术。当前我们所依赖的技术应用方式，正进一步加剧了现有的不平等现象，并加深了已经非常深的数字鸿沟。这是数字合作必须要解决的问题，不仅要致力于帮助人与人之间建立联系，而且还要确保每个人可使用的技术是安全的，并保障使用者的人权。

和平、安全、保证人权，这是合乎人类利益、合乎道德的 AI 趋势在全球范围内不断得到发展的核心动力。Fabrizio 认为，这种 AI 趋势十分重要，它凝聚了全球最聪明的人，而联合国会继续支持这一积极的趋势。Fabrizio 表示，北京智源大会上中国专家和学者提出的 AI 伦理和治理原则，给他留下了深刻的印象，特别是呼吁 AI 以人为本，尊重和保护隐私保护政策的普遍人权。此外，Fabrizio 还注意到了会上频频被提起的“和谐”这一概念，他认为“和谐”在中国具有悠久而丰富的历史，在和谐理念的指导下，使用全球一致的方法来开发和使用 AI 技术是十分必要的，这种全球一致的方法可以容纳各方不同的观点，并找到其中的相似点，然后用和谐的心态处理分歧。Fabrizio 认为，这是一种朝着全球一致的目标前进的建设性思路，这种思路既尊重方法和观点的多样性需求，也关注相似之处，并利用它们为共同的普遍目标和共同的人类价值观服务。Fabrizio 指出，这不是理论上的练习。现在，联合国拥有来自不同实体的 160 多种全球 AI 道德规范和治理原则，通过认识到各种规范之间的相似之处并利用它们建立一致性，会发现它们之间的相似点大于差异，所以 Fabrizio 相信即使在复杂的主题（如 AI 伦理和治理）方面，也可以达成有意义的全球数字合作。

长期以来，这种全球数字合作一直通过科学的方法成功进行，这些方法一次又一次地被证明在开发和平变革性技术（如核能）中扮演了关键的角色。为汲取过往的丰富经验，联合国也正在寻求尝试一些合作模式，尤其是在学术界，北京智源人工智能研究院等智囊团在合作时所体现的模式尤其令人鼓舞。

最后，Fabrizio 感谢了北京智源人工智能研究院、阿里巴巴和其他为联合国的工作做出贡献的组织和群体。他还邀请与会的各位专家与联合国共同合作，共同努力执行秘书长的路线图，以共同塑造一个互惠互利的数字合作未来。

北京市科委主任许强在“人工智能伦理、治理与可持续发展”论坛上的致辞

整理：智源社区 杨香草

6 月 24 日下午，在第二届北京智源大会“人工智能伦理、治理与可持续发展”论坛上，北京市科学技术委员会主任许强出席并致辞。下面是致辞全文。

尊敬的法布里齐奥·霍奇希尔德 (Fabrizio Hochschild) 副秘书长，各位专家，女士们、先生们，朋友们:

大家上午好!

非常高兴与大家共同探讨人工智能伦理、治理与可持续发展的话题。首先，我谨代表北京市科委对面向可持续发展的人工智能智库及公益研究计划的发布表示热烈的祝贺。同时，也向一直以来关心和支持北京人工智能发展的各界人士表示衷心的感谢。

一、AI 伦理和治理已经成为全球共识，各方探索落地的机制

科技伦理是创新驱动发展、数字中国建设、数字时代商业竞争的重要保障。“创新——价值——伦理”形成了一个“铁三角”。创新性技术给社会带来了潜在价值的同时，可能还存在着难以预期的风险，并对社会的伦理提出了重大的挑战。发展人工智能技术，赋能经济与社会的同时，应该关注人工智能的社会属性，从社会风险、伦理准则与治理的角度确保人工智能技术和产业的健康、良性的发展。

当前，国际社会正探索建立广泛认可的 AI 伦理原则，推进敏捷灵活的 AI 治理。2019 年 5 月，经合组织 (经济合作与发展组织，OECD) 通过首部人工智能的政府间政策指导方针，确保了人工智能的系统设计符合公正、安全、公平和值得信赖的国际标准。2019 年 6 月，二十国集团 (G20) 部长会议通过了《G20 人工智能原则》，推动建立可信赖的人工智能的国家政策和国际合作。今年 3 月，联合国教科文组织经过会员国的推荐和遴选，任命了 24 名全球 AI 伦理、政策等领域的专家组成了 AI 伦理特别专家组，已经启动编制了人工智能伦理建议书。智源研究院的伦理与安全中心主任曾毅代表我国入选了该国际组织的专家组。2019 年初，我国成立新一代人工智能治理专业委员会，并于 6 月发布《新一代人工智能治理原则——发展负责任的人工智能》，提出人工智能治理框架和行动指南，强调和谐友好、公平公正、包容共享等八项原则。

北京人工智能产业发展快速，拥有人工智能企业一千多家，占我们全国的近三成；拥有人工智能人才四万人，占全国总量的近六成；拥有首个国家新一代人工智能创新发展试验区和七个国家新一代人工智能开放创新平台，正在加快打造具有全球影响力的人工智能科技创新高地。2019 年 5 月，智源研究院牵头发布了《人工智能北京共识》，为规范和引领人工智能健康发展提供了北京方案。一年以来，相关机构积极推动共识落地。智源研究院推动建设人工智能风险与安全的综合沙盒平台，可对人工智能产品以及应用的风险与安全进行综合检测与评价，降低人工智能产品的全生命周期的风险，对于实现人工智能的自律、善治、有序起到了重要的作用。

二、和谐发展是当前人工智能伦理与治理的主旋律，多方协作，实现发展与治理双轮驱动

和谐发展是当前人工智能伦理与治理的主旋律。一方面，为实现通过人工智能增进人类共同福祉这一目标其途径并不唯一，应当始终秉承与实践“和而不同、和舟共济、和合向善”的发展理念；另一方面，应推动人类与技术的和谐共生，避免技术的误用、滥用、恶用对人类权益的伤害，总体愿景应是发展对人类与生态有益的人工智能。

多方协作，实现发展与治理双轮驱动。治理的目的不是阻碍发展，而是保持发展的健康与稳健。为了确保人工智能向对社会有益的方向发展，应采取发展与治理双轮驱动的战略。通过多轮治理方式，实施对人工智能的研发、部署、使用，从自律自查到顶层监管，真正从技术和社会等不同的视角实现人工智能的敏捷治理。

三、推动各类机构履行社会责任，发展面向可持续发展的人工智能

2015 年，联合国通过了可持续发展目标，共 17 大项，涉及社会、经济和环境三方面的发展问题，致力于从 2015–2030 年间，通过协同行动，消除贫困，保护地球，确保人类共享和平与繁荣。

人工智能是推动社会可持续发展的使能技术。据普华永道预测，到下一个十年年底，将人工智能用于环境应用可能为全球经济带来 5.2 万亿美元的贡献，同时，将温室气体排放减少 4%。麦肯锡的研究显示，在能源、先进电子器件、半导体等 19 个行业中，人工智能的引用每年可创造 3.5 万亿到 5.8 万亿美元的潜在价值。为引导并推动面向可持续发展的人工智能的实现，北京市依托智源人工智能研究院，邀请全球顶尖的专家，成立了科学委员会，组成面向可持续发展的人工智能智库平台，并与全球相关科研机构、企业共同推进面向可持续发展的人工智能公益研究计划，面向全球开放研究成果，支撑人工智能作为使能技术，促进全球社会、经济和环境的可持续发展，推进人类命运共同体的构建与实现。

各位专家，各位来宾，朋友们，人工智能伦理与治理工作关乎全社会、全人类的未来。在当前全球化的背景下，一个可持续发展的地球需要各国在 AI 技术、产业、伦理、治理等各个方面加强合作。

今天，面向可持续发展的人工智能公益研究计划的发布只是一个开始，希望面向下一个十年，全球各类 AI 机构能继续秉承以人为本及可持续发展理念，推进人工智能治理的跨学科差异和国际合作，共同携手推进全球人工智能的健康可持续发展。谢谢大家。

Danit Gal: 人工智能伦理全球合作：东亚人工智能伦理的视角

整理：智源社区 罗丽

联合国秘书长技术顾问，剑桥大学智慧未来研究中心兼职研究员 Danit Gal 的报告主题是《Global cooperation on the Ethics of AI: A look at East Asia》（人工智能伦理全球合作：东亚人工智能伦理的视角）的特邀报告。

报告中，Danit Gal 从东亚 AI 伦理发展、欧盟和美国 AI 伦理发展以及 AI 理论发展全球合作三个方面，介绍了东亚人工智能在全球背景下的发展现状，东亚 AI 伦理发展对全球 AI 理论发展所做出的贡献以及 AI 理论全球合作的必要性和发展方向。

在上海市科学学研究所发表的《全球人工智能治理年度观察 2019》中，Danit Gal 论述了 AI 发展由原理转向应用的必要性，Danit Gal 的研究发现，在不同的国家和地区，AI 的应用方式也会有所不同，她提出只有了解了这些差异，才能对 AI 伦理学进行有意义的全球对话。但在撰写报告时，Danit Gal 产生了另外一个问题，理解应用中的差异是否足以在 AI 伦理方面建立有意义的全球合作。本次报告中，Danit Gal 更进一步地介绍了 AI 理论在东亚、欧盟以及美国的发展现状，并提出人工智能伦理全球合作所面临的问题及发展方向。

一、东亚 AI 理论发展

东亚人工智能理论（Perspectives and Approaches in AI Ethics: East Asia）的研究是基于 Danit Gal 为《牛津人工智能伦理学手册》（2020 年）所进行的研究。

Danit Gal 表示，在韩国，人工智能与人类互动被描述为 Shared Social Responsibility，即共享的社会责任。早在 2007 年，韩国就建立了明确的人机交互等级制度，以避免破坏个人利益和集体利益之间的微妙平衡。以人为中心的人工智能是人类和人工智能之间相互作用的结果，韩国政府强调需要避免反社会发展，因为 AI 会转移和分散人与人之间的交流和互动，以一种损害人类的方式的方向来取代人类。为了避免反社会发展，韩国政府要求用户、开发人员和公司对 AI 伦理的发展负责。

在中国，人工智能与人类合作被描述为 Human – AI Harmony，即人类与 AI 和谐共生。和谐的概念在中国传统文化中占有特殊地位，根据和谐理念，中国发布了中国 AI 伦理原则，旨在指导 AI 与人类的和谐发展，中国的“和谐人工智能原则”为人类和人工智能提出了共同的原则，即，以共同的命运发展战略性未来。Danit Gal 认为，在中国的和谐人工智能环境下，中国的 AI 政策、AI 教育、AI 培训等领域的发展将具有广阔的探索空间。

接下来，Danit Gal 介绍了日本的 AI 治理。在日本，政府提出了 Co-evolution and Co-existence（共同进化与共存）的 AI 发展理念。日本人工智能学会提出的 Society 5.0 旨在创造人类和智能机器在一种完全技术支持的社会中共同进化和发展，并提出想要“成为社会准成员”的人工智能所必须遵循的道德原则。日本认为，将 AI 技术融入社会能取得经济效益，也是减轻国家孤立、应对超级老龄化社会带来的挑战的一种应对手段。

二、欧盟和美国的 AI 理论发展

Danit Gal 介绍，欧盟的 EU Humane AI project 致力于通过了解我们人类、我们的社会以及我们周围的世界来

增强和授权全人类 AI 系统。欧盟委员会提出，人工智能技术的开发应以人为中心，因此值得公众参与……人工智能应用应赋予公民权利并尊重其基本权利，人工智能应致力于提高人们的能力，而不是取代人类等人工智能发展理念，这些发展理念旨在创造一个清晰的人与机器的等级制度。在 MIT Human–Centered AI Collective 中，学者们提出“AI 系统必须通过向人类学习，不断改进”、“创造有效且充实的人机交互体验”等理论。人工智能如何提供这些服务，同时创造一种友好和令人满意的人机交互体制？如果从不同的文化角度来看，答案可能不同。在 Stanford Human centered AI Institute，研究者表示如果人工智能要满足人类的集体需求，那么它就必须对人类身体上、智力上和情感上的动因进行理解，设计能够理解人类语言、感觉、意图、行为和具备区分细微差别和多维互动能力的机器智能，这些都是至关重要。

那么，不同国家和地区的人工智能与人类关系的不同之处在于什么？

人类与人工智能之间的兼容性语言与英语相似，但重点却不同，比如，一些西方话语着重于 AI 对人类的理解，而某些东亚语言则着重于人类与 AI 的相互理解。这种差异可能对 AI 本身的发展并不是那么重要，但对采用和使用 AI 的广大公民而言，这些差异是至关重要的。

在《牛津人工智能伦理学手册》的研究中，Danit Gal 提出，受欢迎的包含技术的科技动画文化具有更高的社会采用率和更快、更深入的社会嵌入能力。但是具有科技动画文化的人工智能应用可能会与其他非技术性动画文化产生冲突。

图 1：科技动画文化应用实例

上图为具有科技动画文化的人工智能理论的应用实例。图 1 是一个在中国寺院中成长的长袍僧人，通过使用 AI 设备来传播信息，图 (2) 是一个被指定提供的机器人 Sarah，机器人 Sarah 能够以较低的成本来哀悼死者。图 (3) 是来自中国和日本的虚拟偶像，其听众来自全世界并且具有庞大的粉丝群。图 (4) 是日本公司 Gatebok 所设计的虚拟妻子，在没有出现之前便被预售而空。这些实例意味着，包含技术的科技动画文化可能具有更高的社会采用率和更快更深的社会嵌入能力，也可能会与其他非技术性动画文化产生冲突，AI 技术的不同应用对现实生活具有完全不同的影响。

微软的成功，是了解区域差异并从中受益的有力示例。一切都是从 Teams 开始的，它是一个聊天框，通过 Twitter 被发布到外界，但在即时通讯信息发布 24 小时后，仍毫无音讯，刚开始，它被认为是一个失败的实验，并且得到了很多负面媒体的报道。之后，设在中国的微软亚洲研究小组承担了这个项目，并把它变成了迄今为止世界上最成功的社交聊天箱，目前为止，它已经拥有 6.6 亿的用户互动数，在某种程度上，被认为是一个交互奇迹。研究小组创造的日本用户聊天机器人 Rinna，在某种程度上得以流行，之后，他们试图回到西方用户，通过创造与政策想近的话题来吸引用户，实验表明，之后确实有几百个、几千个的互动产生，并不断取得成功。所以，得到的一个有趣的实验结论是，机器人在东亚的流行和成功，是因为在那里的人们愿意与 AI 进行有意义的交流和对话。

三、AI 伦理全球合作

Danit Gal 表示，如果微软没有成功，我们如何真正追求人工智能伦理的全球合作？ Danit Gal 以微软的成功为例，讨论了实现人工智能伦理的全球合作需要解决的三个障碍：语境、交流、合作。

(一) 语境

在西亚和东亚的用户与开发人员之间，人与 AI 之间的交互通常在语境上有所不同，这种语境对 AI 的接受和使用起着重要作用，会产生不同的开发和使用轨迹，如果不能正确传达用户意见、发展和使用上的差异，将无法解决问题并会阻碍合作。

(二) 交流

语境不同将增加交流难度，需要考虑语境在文化和社会结构中的嵌入程度。语境不同导致有关 AI 伦理全球合作中的一个已知困境，即我们使用相同的语言（英语）来传达相似的原理，并且（希望）认识到有不同的实现方式。 但为什么？如果无法有效地沟通分歧，将导致我们无法有效地沟通如何去解决分歧。

(三) 竞争

人类使 AI 成为一项竞争性技术，因此缺乏交流的动力。AI 的货币化和政策化源于对比性叙述和缺乏的交流沟通，以致出现“ AI 军备竞赛”、“ AI 统治 = 世界统治”之类的呼声。因此，如果没有有意义的动机来处理和解决国家和区域之间的差异，我们就永远无法解决这些差异，所以只能选择继续竞争。

那么，如何来解决这些问题呢？ Danit Gal 表示，应该从三个方面入手:

(1) 可以通过探索、理解和交流差异来促进有效的、非竞争领域的合作。
(2) 创造有意义的激励机制，比如通过可持速发展和使用开源技术等方式创造一个没有竞争的合作环境，使每一个人都能从中受益。
(3) 一起致力于 AI 伦理及其他方面的全球合作。

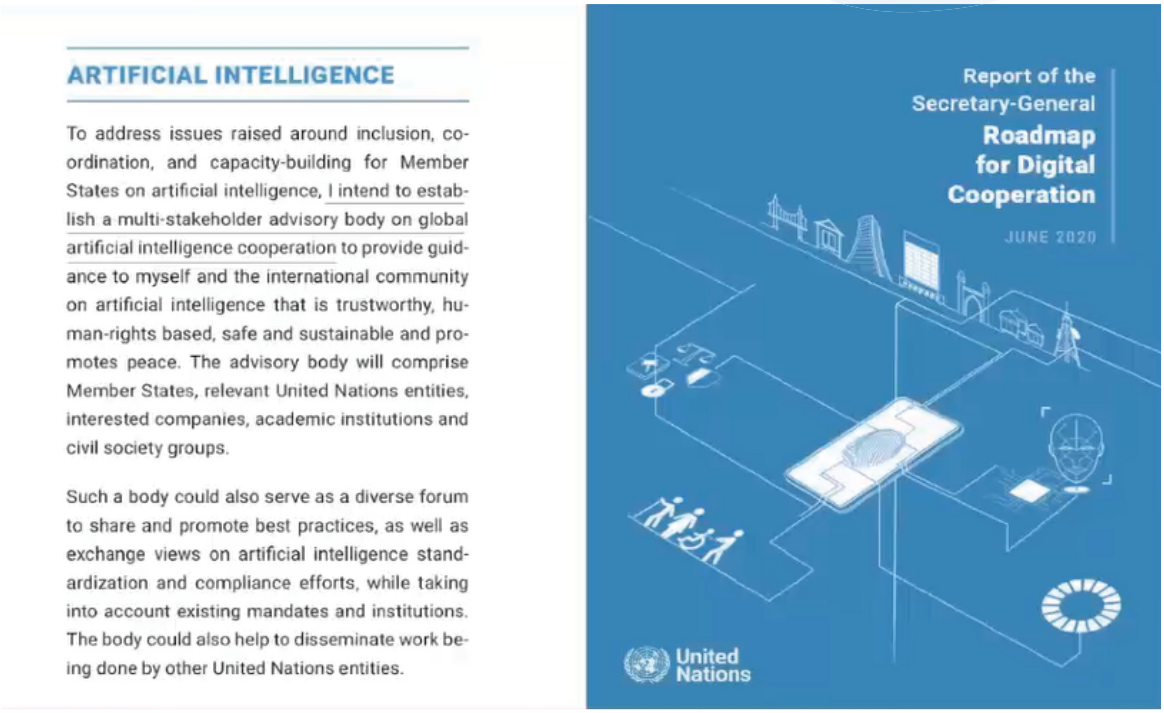


图 2：联合国数字合作路线图的报告

最后 Danit Gal 介绍了联合国最近发布的数字合作路线图，并表示，联合国秘书长已宣布将建立全球人工智能合作多方利益原则，建立世界上最具包容性和信息量的机构，并欢迎所有有志之士的加入！

联合国秘书长数字合作高级别小组前执行主任 Amandeep Singh Gill：通过伦理实现信任——解锁 AI 的可持续发展使用的关键因素

编辑：智源社区 杨香草

Amandeep Singh Gill 的演讲主题是《通过伦理实现信任：解锁 AI 的可持续发展使用的关键因素》

Amandeep Singh Gill（阿曼迪普·辛格·吉尔）拥有多重身份，他是国际数字健康与人工智能研究合作中心（International Digital health & Artificial Intelligence Research Collaborative，简写为 I-DAIR）的项目主任。他是联合国秘书长数字合作高级别小组前执行主任，该特别工作组联合主席还包括比尔盖茨的妻子梅琳达·盖茨（Melinda Gates）和阿里巴巴创始人马云。Amandeep 曾任印度大使及日内瓦裁军谈判会议常驻代表，印度外交部裁军与国际安全司司长（2013–2016 年），且于 2017 年协助成立印度经济转型人工智能工作组。

在演讲报告中，Amandeep 认为 AI 应用的关键是通过制定伦理原则来建立人类与 AI 之间信任的基本观点。他描述了 AI 的发展现状和其面对的机遇与挑战，强调了 AI 治理的重要性，提出了 AI 治理需要遵循和践行的原则性方法。

图 1：Amandeep 的演讲主题：通过伦理实现信任：解锁 AI 的可持续发展使用的关键因素

一、AI 技术和 AI 治理面临的机遇与挑战

Amandeep 认为，当人们意识到 AI 技术可能有助于实现可持续发展目标（Sustainable Development Goals，SDGs）之后言，AI 技术的机遇存在于：

1. 改变实现方式。AI 技术可以提供横向思维并节约成本来重塑 SDGs 的实现方式。例如，重塑 SDGs 中的目标 3，即健康与福祉 (Health and Well-being)[1]。
2. 干预措施。AI 技术能将机器学习、大量数据集和计算能力融合在一起，为许多领域提供果断的干预措施。比如在健康领域，AI 技术能够精准预防并预测健康情况，所以流行病问卷调查在新冠肺炎疫情期间受到重视。
3. 数字机遇。受新冠肺炎疫情影响，注意力转移到了弹性投资上，数字机遇被看好。数字机遇不仅仅是炒作，其前景可观。

机遇总是伴随着挑战。Amandeep 将 AI 技术面临的挑战归纳为三点:

1. 缺失数据 (Missing Data)。尤其是在中低收入国家，缺少发展数据，没有数据基准。
2. 滥用数据 (Misuse)。面对数据滥用，公众会关心数据安全、数据所有权、数据隐私和知情同意等方面的内容。
3. 漏用数据 (Missed Use)。由于数据质量差，数据集分散，缺乏互用性，对于数据持有零和 [4] 观点和态度也会导致信任不足。

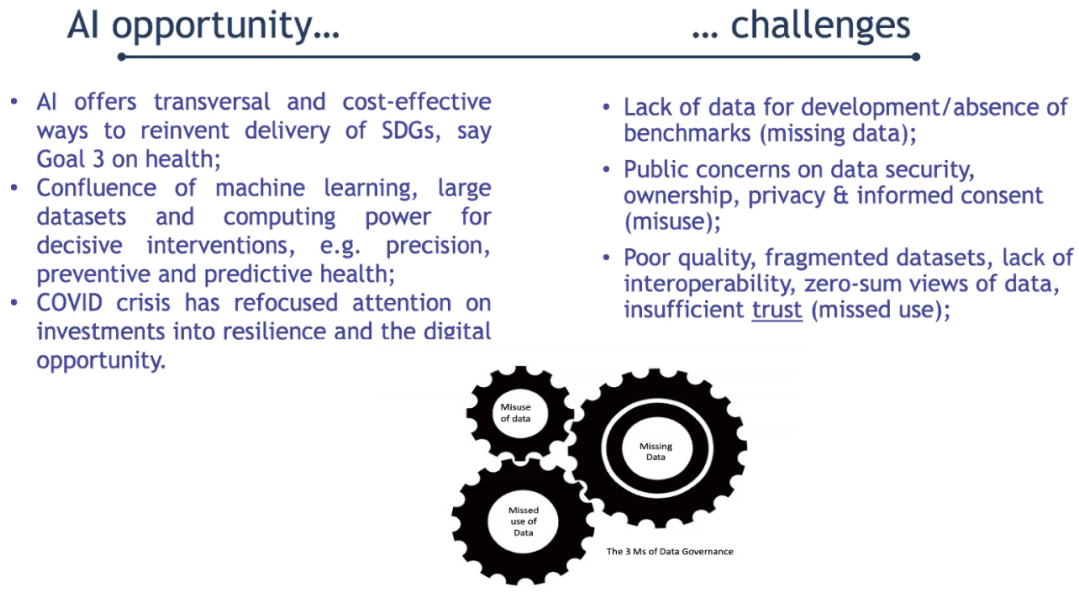


图 2：AI 技术面临的机遇与挑战

AI 治理也面临着诸多具体的挑战:

1. 治理方法不统一。在科学和数据层面，民族主义世风日盛。
2. 炒作。AI 模型过度拟合 (Over Fitting) 或欠拟合 (Under Fitting)，从而导致领导人陷入困惑，判断失准，不知道是要立即治理 AI 还是滞后处理。
3. 对待数据或技术的集中式方法冲击了分散式方法。分散式方法尊重本地数据或技术所有权，允许亚洲和非洲等地新兴区域创新扮演好自己的角色。
4. 国际组织和领导人常常终止让人失望的创新治理，并过早宣布达成共识。
5. 缺乏治理能力，缺少基本的对健康领域使用数据和 AI 技术的跨域理解能力。

6. 太多的政策原则和抽象性太强，施行拟定的治理方案证据有限。我们需要具体的施行办法，并用证据来支持。

图 3：AI 治理面临的挑战

二、AI 技术、价值观和原则及 AI 伦理治理的辩证关系

技术从来都不能脱离一定的价值观，也做不到中立似流水，隐去特性。事实上，技术反映人类的价值观和偏好层次，而价值观和原则 (Values and Principles, Vs & Ps) 有助于创建伦理文化，构建价值观驱动的技术发展和技术治理。这有助于形成价值观—伦理行为—信任技术和技术治理的良性循环。

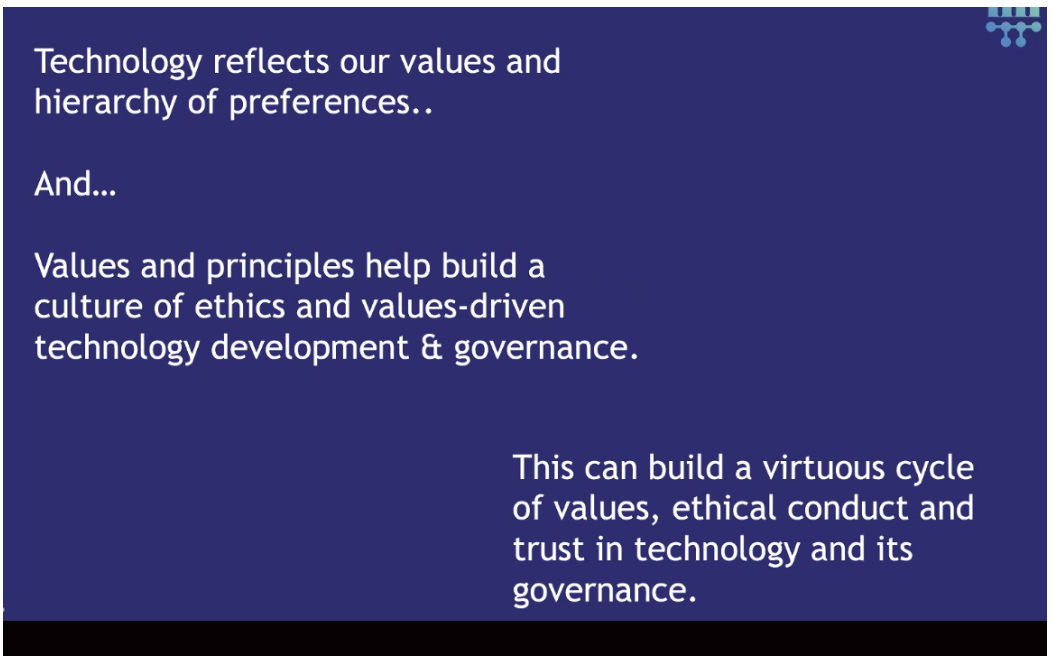


图 4：技术，价值观和伦理治理的关系说明

笔者用图 5 来表达以上概念之间的辩证关系，方便读者有一个直观的了解:

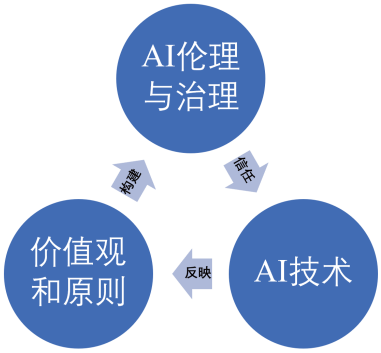


图 5：AI 技术、价值观和原则及 AI 伦理治理的辩证关系图示（笔者注）

三、价值观和原则：从概念到行动

厘清这些关系之后，要从抽象概念落到实处。那么
AI 治理政策和行动落实的切入点又在哪里呢？ Amandeep 认为:

1. 价值观和原则能指导政策制定和实施;
2. 价值观和原则能将国际准则通过非正式渠道传播到各国;
3. 价值观和原则使得治理范围扩展到技术发展周期的早期和模糊不清的部分;
4. 价值观和原则能够跨文化跨国界灵活运用。
5. 在找到从抽象走向具体的切入点之后，在落实具体措施之前，还有一些因素要考虑到:
6. 价值观和原则要出现在每个使用语境中，并与治理成果显著相关;
7. 价值观和原则要反映不同文化或国家背景;
8. 价值观和原则要做到透明化，避免被潜在的商业利益或政策利益操纵 (“伦理道德洗脑”，‘ethics wash’);
9. 要避免对价值观和原则的滥用，尤其是在现有的数字治理不对称的背景下，避免 AI 治理过程中的集中式决策行为 (“治理中的网络效应”，‘network effects on gov’ [2])。

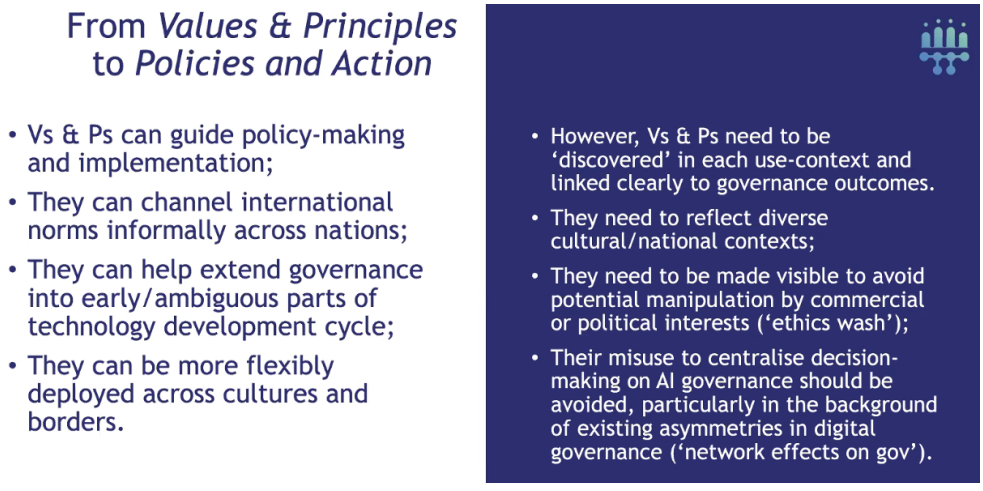


图 6：从价值观和原则到政策与行动

最后，Amandeep 总结了五种实用的伦理价值观来指导 AI 技术应用全程有序稳定发展：

1. 修复世界 (Tikkun Ulam)[3]。当今世界有很多期待解决的难题，SDGs 就是致力于解决这些难题，而这也是善用 AI 技术的重点方向。
2. 人类繁荣 (Human Flourishing)。追求健康，幸福和繁荣，实现经济向好发展，这是 AI 技术要发力的方向，同时注意平衡社会发展。
3. 和谐 (Harmony)。“和谐”的概念源自中国，AI 技术不应该是用来制造社会内部或社会之间的冲突，破坏人权和自由，而是允许人类一起解决问题，实现繁荣发展。
4. 人道 (Ubuntu)[4]。这是非洲的哲学价值观，意为你决定我即是我，不管我做什么，都是在修复每个人。所以这种价值观倡导一种责任感，倡导形成社区，倡导在 AI 技术层面建立人类社会。
5. 制感 (Pratyāhāra)[5]。这个价值观来源于印度，意为内省，但并非与外面世界隔离，意为我不是我的感官意识，我可以向内探索找到我的本质。在使用 AI 技术时，也要融会贯通这种价值观，多关注技术所服务的本质。

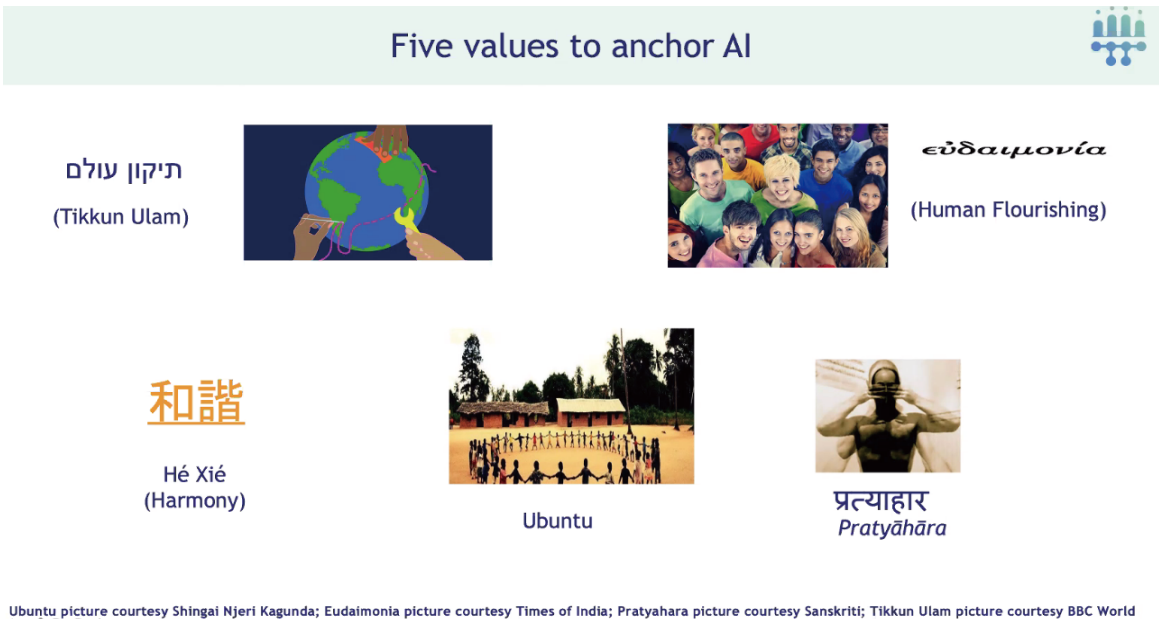


图 7：Amandeep 总结的五种价值观

参考注释：

[1] 据百度百科，联合国可持续发展目标是一系列新的发展目标，将在千年发展目标到期之后继续指导 2015–2030 年的全球发展工作。2015 年 9 月 25 日，联合国可持续发展峰会在纽约总部召开，联合国 193 个成员国在峰会上正式通过 17 个可持续发展目标。健康与福祉是第三个具体目标。可持续发展目标旨在从 2015 年到 2030 年间以综合方式彻底解决社会、经济和环境三个维度的发展问题，转向可持续发展道路。

[2] 治理中的网络效应 (network effects on governance)：网络效应，也称网络外部性或需求方规模经济，由以色列经济学家奥兹·夏伊 (Oz Shy) 在《网络产业经济学》(The Economics of Network Industries) 中提出。在具有网络效应的产业中，“先下手为强”(first–mover advantage) 和“赢家通吃”(winner–takes–all) 是市场竞争的重要特征。由此可知，治理中的网络效应指的是价值观和原则强势国家的集中式治理，会使得治理权力失衡，并带来一系列的相应影响。

[3] 修复世界 (Tikkun Ulam)：意为 fixing the world，犹太教中的概念，通常被解释为渴望表现出建设性和有益的行为。

[4] 人道 (Ubuntu)：非洲南部祖鲁语或豪萨语，意为人性，“我的存在是因为大家的存在”，是非洲传一种传统的价值观。南非总统曼德拉认为它包含了尊重、互助、分享、交流、关怀、信任、无私等众多内涵，是一种生活方式，提倡宽容和同情他人。

[5] 制感 (Pratyāhāra)：是印度《瑜伽经》八支中的第五支，强调精神从感觉和外部事物的奴役中解脱出来，是指感觉消失，控制内心，也称调心。

全球 AI 伦理协会发起人 Christoph Lütge：人工智能和可持续发展

编辑：智源社区 杨香草

在本次智源大会“人工智能伦理、治理与客持续发展”论坛中，全球 AI 伦理协会发起人 Christoph Lütge 做了主题为《人工智能和可持续发展》的报告。Christoph 的报告围绕着“AI 技术为实现可持续发展目标大有可为”进行展开。

Christoph Lütge（克里斯托夫·卢奇）是德国哲学家和经济学家，慕尼黑工业大学商业伦理学教授兼人工智能伦理学研究所所长，与东京大学，纽约大学和剑桥大学等大学的科学家共同创立了全球人工智能伦理学联盟。他是全球 AI 伦理协会发起者，以其在商业伦理，人工智能伦理，实验伦理和政策哲学方面的工作而闻名。

在该演讲中，Christoph 重点介绍了其所任职的人工智能伦理中心的研究情况，人工智能技术为可持续发展目标服务的可能性和研究领域，以及人工智能伦理原则的内容。这三部分的内容都很精彩，尤其是在伦理原则部分表达出的“以人为本”和“信任是使用的前提”的中心思想是很值得期待的闪耀着智慧光芒的见解。

以下为 Christoph 演讲全文整理，以供交流学习。

图 1：Christoph 演讲主题：人工智能和可持续发展

一、TUM-IEAI 研究现状简介

Christoph 在进入主题之前，先就慕尼黑工业大学人工智能伦理中心进行了介绍。多年来，慕尼黑工业大学 (Technische Universität München, Technical University of Munich, TUM) 是科学、技术和社会的互动研究方面的先驱，立足于“以人为本的工学”。人工智能伦理中心 (The Institute for Ethics in Artificial Intelligence, IEAI)[1] 成立于 2019 年，隶属于慕尼黑社会技术中心 (Munich Center for Technology in Society, MCTS)。

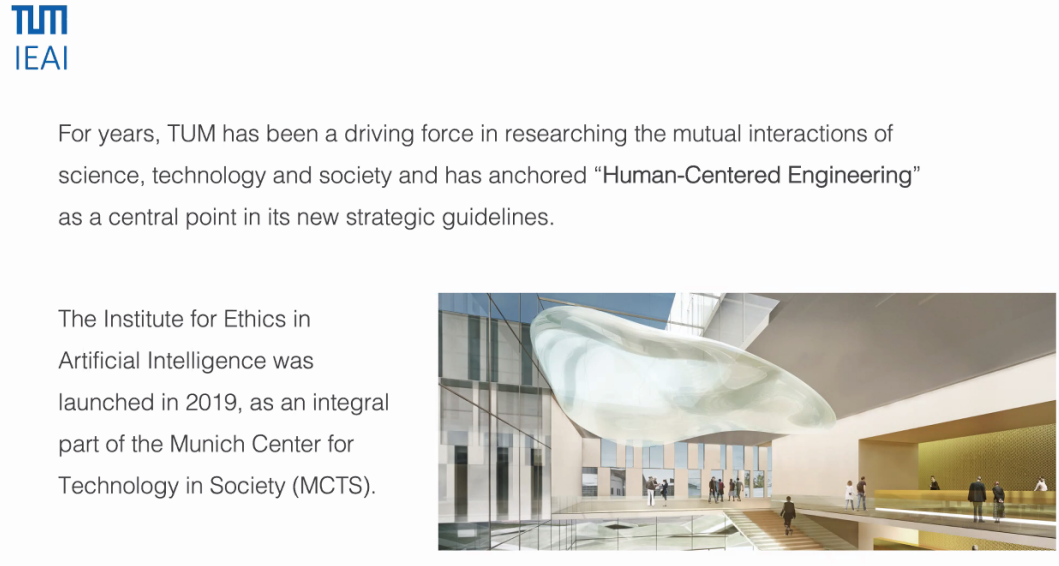


图 2：慕尼黑工业大学人工智能伦理中心 (TUM–IEAI) 一览

IEAI 的使命不仅是探索 AI 技术更多的可能性，也是确保让更多的人享受 AI 技术带来的红利。IEAI 与自然科学和人文社会科学方向都有合作，构建结合 AI 技术的未来合作指南。IEAI 合作的方向包括法律、治理 / 政策、商业、伦理、数学、医学、计算机科学、工程学等，在公平和跨学科原则指导下，建立全球合作，实现全社会的 AI 技术伦理发展与实践。

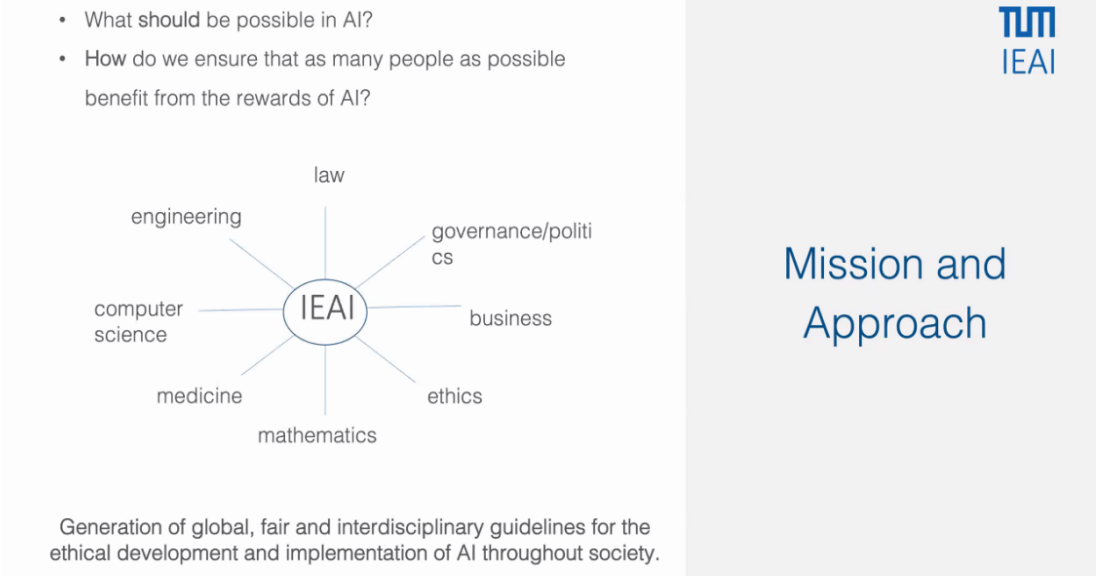


图 3：IEAI 的使命与技术方向

IEAI 的业务内容，包括:

1. 为工业界和学术界之间的合作提供平台，还包括研究项目、工作坊和会议等，合作网络不断扩大。

2. 支持并协调在慕尼黑工业大学进行跨学科研究，研究包括 IEAI 内部资助项目及其他外部资助项目。

接下来，Christoph 介绍了 IEAI 当前的研究重点。IEAI 关注跨学科方向的研究，合作单位包括各个学科部门，工科合作部门有电子与计算机工程系、信息学系、机械工程系、航空航天和大地测量系等，还合作有医学部，慕尼黑社会技术中心、数学系、数据科学研究所等理科部门，文科合作部门包括 TUM 的教育学院、治理学院、管理学院等。IEAI 研究小组有：AI 与未来的工作，AI、移动出行与安全，AI、选择与自主性，AI 在医疗保健方面的应用，AI 与线上行为，AI、治理与规约，AI 与可持续发展等。下图 4 和图 5 是 IEAI 跨学科研究项目的部分汇总。

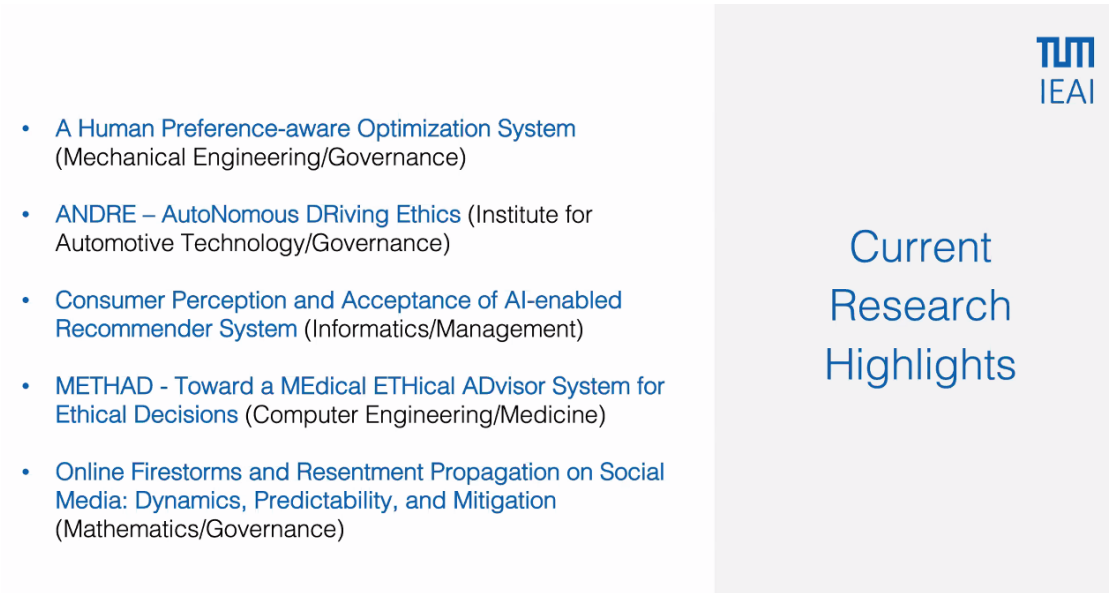


图 4：IEAI 跨学科研究项目一览（一）

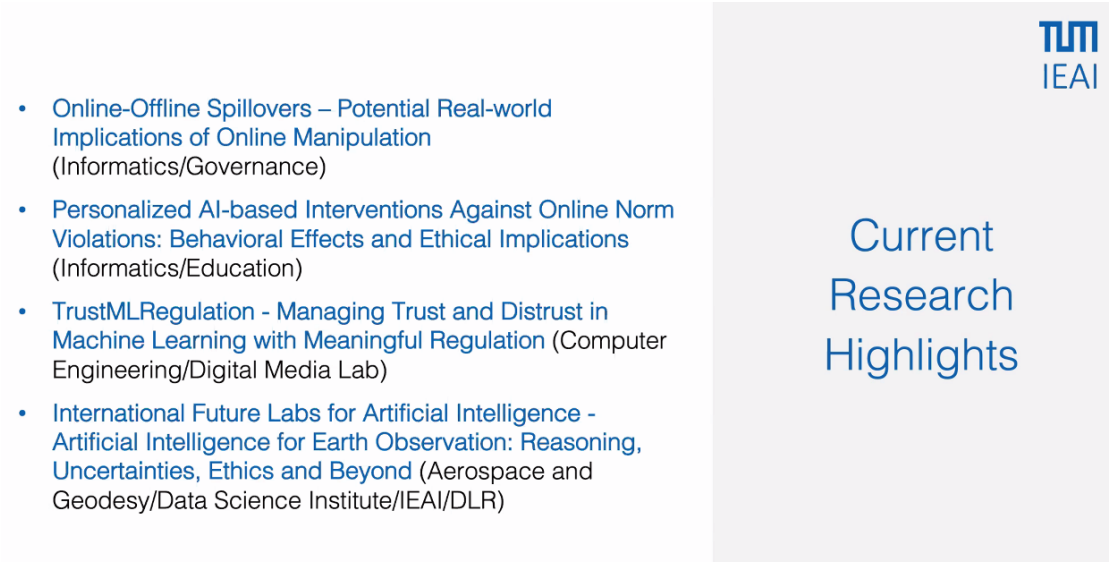


图 5：IEAI 跨学科研究项目一览（二）

在介绍 IEAI 之后，Christoph 紧接着提到了国际人工智能伦理联盟 (Global AI Ethics Consortium, GAIEC)[2]。GAIEC 的目标是协调和倡导独立学术研究，着眼于设计基于 AI 的技术的应用框架和落实伦理发展指南。下图 6 是 GAIEC 的发起单位和代表列表，包括 Christoph 本人所在的 IEAI 以及本场论坛主持人曾毅所在的中国北京智源人工智能研究院伦理与安全研究中心。值得一提的是，包括上述两个机构在内，该联盟中的 9 家机构联合发起了“全球人工智能伦理与抗击新冠疫情联盟”[3]，为抗击疫情提供帮助。

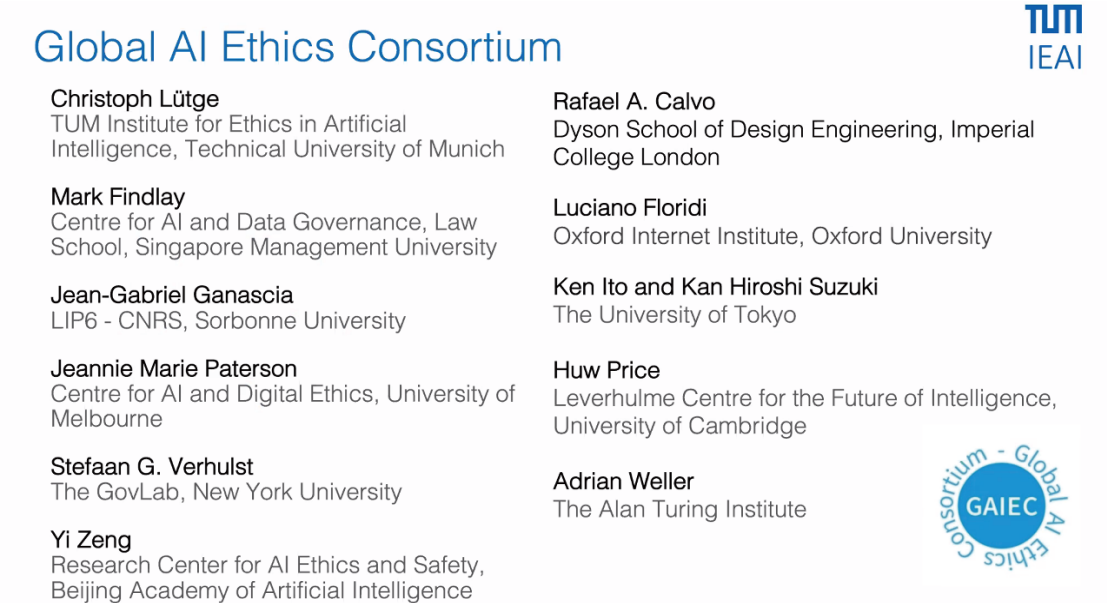


图 6：GAIEC 的发起单位列表

紧接着，Christoph 的演讲进入 AI 与可持续发展的主题，该主题涉及范围很广，Christoph 介绍了 IEAI 利用 AI 技术进行的相关研究项目，即利用 AI 技术观测地球 (AI4EO)[4]。AI4EO 实验室由慕尼黑工业大学和德国航空航天中心 (The German Aerospace Center, DLR) 合作成立，研究期限为 3 年，从 2020 年 5 月至 2023 年 5 月。

AI4EO 汇集了 9 个国家 20 个知名国际组织中的 27 名高级别科学家，来共同应对三个基本挑战：推理 (Reasoning)，不确定性 (Uncertainties) 和伦理 (Ethics)。该项目致力于解决数据保护 / 隐私问题，遵守数据可携权 (Data Portability)[5]，确保在数据收集层面的公平和平等，以及确保数据使用和传播的公平和平等，并将其纳入伦理治理范畴。

Christoph 通过介绍该项目进行的一个关注印度某城市的贫民区的专题研究，展示了 AI 应用于此领域的前景。通过 AI4EO 与水文经济多智能体模型及利益相关者知识的耦合，该项目计划创造一个规划工具，用来捕获在贫民区普遍存在的复杂供水系统的相关反馈动态。因此，AI4EO 能够为形成一个全新的基础设施规划范式打下基础，并最终以少花钱的方式，扩展贫民区的公共规划能力，从而改善地区供水现状。

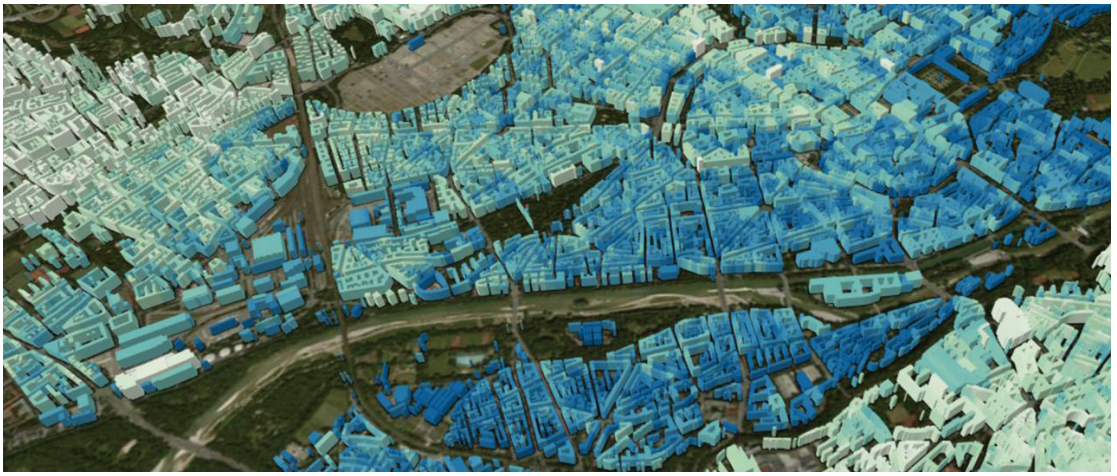


图 7：城市的全球 3D 模型的一部分 [6]

AI4Slums 专题由慕尼黑工业大学、亥姆霍兹环境研究中心 (UFZ)、德国航空航天中心、奥地利发展研究基金会 (ÖFSE) 及斯坦福大学合作开展，预估将于 2021 年启动，在该专题中，IEAI 扮演着提供技术伦理指导的角色。Christoph 认为这一角色如果缺失，危害性会体现在两方面：一，如果输入数据存在的潜在偏见，会导致一系列不准确的输出结果或信息传送给利益相关者；二，从 AI4EO 获取的贫民区及其居民的有关信息，会被不同的使用者滥用，例如，用来压制贫民区的扩大，剥夺贫民区现有居民和潜在居民的限定财产权 (qualified interest)[7] 。

二、AI 技术应用于 SDGs 的领域

联合国可持续发展目标 (Sustainable Development Goals, SDGs) 是联合国制定的 17 个全球发展目标，在千年发展目标 [8] 到期之际继续指导 2015–2030 年的全球发展工作。Christoph 将 17 个目标归入社会，经济和环境三个方面，而 AI 技术可以在这三个方面为 SDGs 服务。

图 8：联合国可持续发展的 17 个目标

与社会相关的 SDGs 包括 SDG1 消除贫困，SDG2 消除饥饿，SDG3 良好健康与福祉，SDG4 优质教育，SDG5 性别平等，SDG6 清洁饮水与卫生设施，SDG7 廉价和清洁能源，SDG11 可持续城市和社区，SDG16 和平、正义与强大机构； 与经济相关的 SDGs 包括 SDG8 体面工作和经济增长，SDG9 工业、创新和基础设施，SDG10 缩小不平等差距，SDG12 负责任的消费和产品，SDG17 促进目标实现的伙伴关系； 与环境有关的 SDGs 有 SDG13 气候行动，SDG14 海洋环境，SDG15 陆地生态。

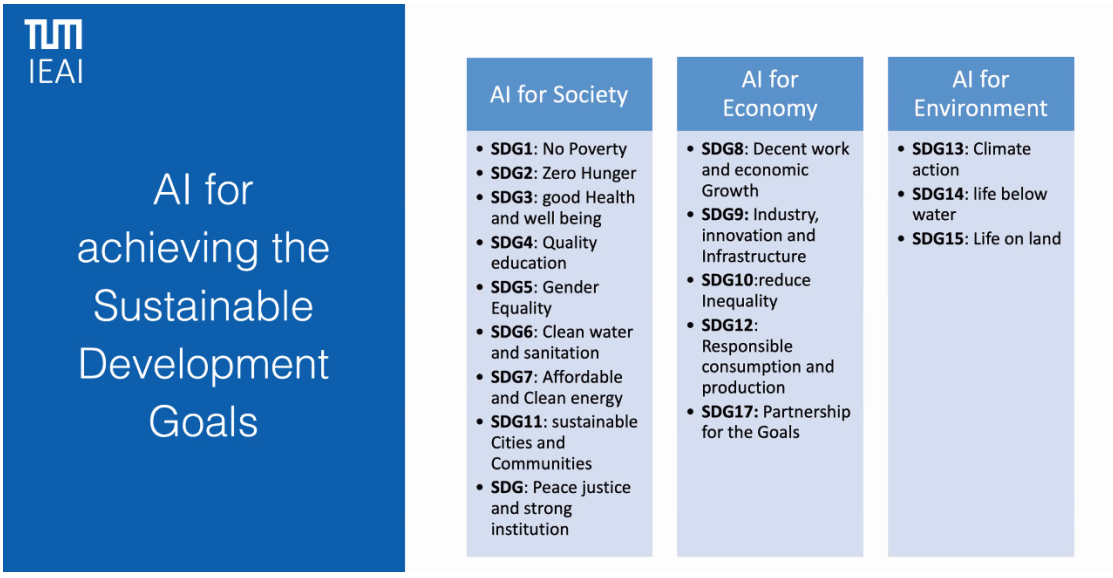


图 9：联合国可持续发展目标与 AI 技术的结合

SDGs 是基于现实问题提出来的，21 世纪的世界面临着许多挑战，比如人口问题，到 2050 年，地球居住人口将达到 95 亿； 比如资源与环境问题，我们如何在为人类提供清洁的饮用水、教育及可持续发展能源的同时，保护好环境?

那么 AI 技术能帮助人类解决这些问题吗？ AI 技术作为工具，可以有效解决我们面临的饥饿、饥荒、流行病等关键问题，而且 AI 技术让缩短寻找解决办法的时间成为可能。利用 AI 技术实现可持续发展目标 (Sustainable Development Goals, SDGs) 的应用前景广阔。

Christoph 挑选了几个 AI 可以大展身手的领域进行了介绍 :

- 利用 AI 技术来保护濒危物种 (AI to help Protecting Endangered Species)。现在的动物保护主义者利用红外相机等设备来确认物种栖息地，然而这也存在一定的问题，红外相机允许收集数百万张照片的数据，这也就意味着处理数据耗费时间长，关键信息可能会遗漏或失效。运用 AI 技术之后，我们就拥有了世界上最大的红外相机数据库，深度学习模型能为动物保护技术提供最优和最新数据，加速收集、分析及在全球分享野生动物相关数据。
- AI 技术改革移动出行方式 (AI to Revolutionize Mobility)。全球二氧化碳排放量的四分之一归于运输行业使用能源排出的废气。面对此种现状，大规模使用自动驾驶汽车有巨大的减排潜能。在互联移动领域进行创新，创造智能出行工具，建设智慧型基础设施。这些措施可以有效减少交通拥堵，建设高效和环境友好型运输系统。

- AI 技术可以救助生命 (Life-Saving AI Technology)。AI 技术能够预测食物短缺。2018 年，联合国秘书长、难民署高级专员古特瑞斯 (António Guterres) 建立了饥荒行动机制 (Famine Action Mechanism, FAM)。FAM 是第一个量化模型处理机制，利用算法实时计算食物安全性。由此看来，AI 系统能够通过检测不同变量之间的相关性来预测食物短缺风险。AI 技术使人类具有在早期预警系统开始的同时提前准备资金以预防食物安全危机。
- AI 技术有助于提高粮食产量 (AI for Efficient Food Production)。AI 系统能帮助农民实时分析天气状况、温度、用水情况及土壤环境，从而提高粮食质量，提供准确作物收获时间信息。这种精准农业 (precision agriculture) 利用 AI 技术检测植物病虫害情况和农作物营养不良情况。AI 技术创造的季节预报模型提高了农业精准度，提高了生产能力，该模型能够提前预测好几个月的天气情况，有助于农民更好地安排农活。
- AI 技术可用于循环系统 (AI for Recycling)。AI 分类机器人 (比如芬兰泽恩机器人公司，ZenRobotics Ltd.) 可以做到智能机器将垃圾分类。机器人使得循环处理更加高效、准确并获利。除此以外，机器人也能减少环境污染，甚至减小环卫工人每天在回收厂面对的健康危害。

三、AI 技术面临的挑战及 AI 伦理原则

通过以上应用前景可以真切地感受到 AI 技术能够服务于人类，但同时也面临着技术挑战:

- 依赖于技术系统的准确性，就要承担技术错误的风险，比如远程医疗 (Telemedicine)。同时也会有失去自主决策权的危险。
- 针对网络攻击变得更加脆弱不堪。
- 侵犯隐私和数据滥用的风险。欧盟通用数据保护条例 (GDPR) 可以有效应对此风险，但又需要进行条例调整来抗击疫情及其他流行病。要注意的是，不同的人对此问题有不同的看法。
- 数字素养 (Digital Literacy)[9]。现在通常缺少对于 AI 技术的数字素养教育。在医疗健康领域，医生不信任 AI 系统。

除了技术挑战以外，还有关于 AI 伦理方面的担忧。AI 技术出于什么目的使用什么数据在多大程度上是透明的 (Transparent)，公司能够提供的信息是可审核的 (Auditable)，数据使用的全程流向是可解释的 (Explainable and Interpretable) 以及数据使用是公平的 (Fair)。

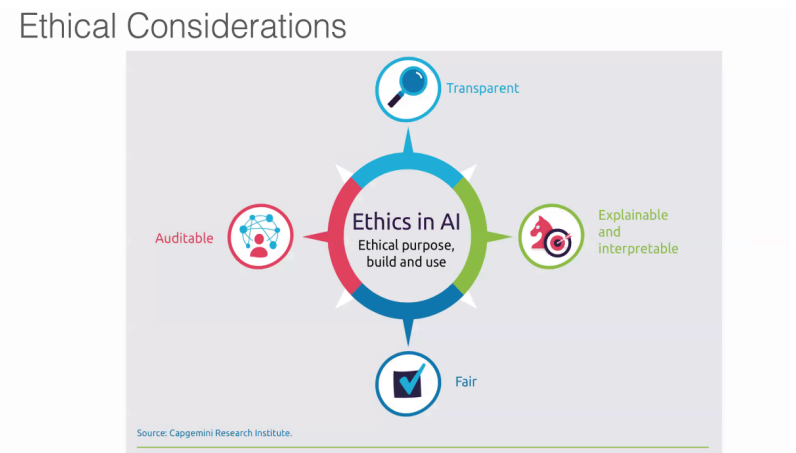


图 10：AI 技术伦理方面的担忧

Christoph 提出了 AI 伦理的一个中心和五个原则。一个中心指的是 AI 技术要以人为本 (Human centered approach)。五个原则分别是:

1. **行善原则 (Beneficence)**。具体可解读为提高福祉，维护尊严，地球存续。
2. **不伤害原则 (Non-maleficence)**。具体可解读为保护隐私，数据安全，“能力警告”(capability caution，指的是未来 AI 技术能力的上限预警)。
3. **自治原则 (Autonomy)**。具体可解读为在人类继续持有的决策权与人类委托给 AI 技术的决策权之间取得平衡。
4. **公正原则 (Justice)**。具体可解读为创造共享可享的利益，保持团结。
5. **可解释原则 (Explicability)**。具体可解读为要确保其他原则的可理解性 (Intelligibility) 和可信性 (Accountability)。

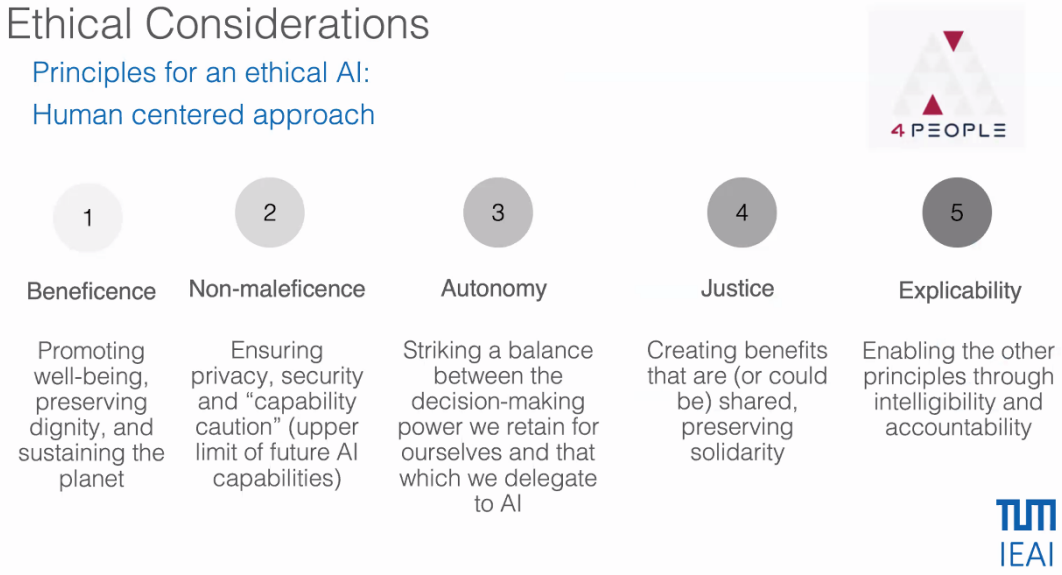


图 11：AI 伦理的一个中心和五个原则

Christoph 随后列举了几个相关例子 :

- **经济合作与发展组织 (Organization for Economic Co-operation and Development, OECD)**[10]。OECD 有三十多个成员国，其奉行的 AI 原则 (2019 年 5 月正式通过 OECD Principles on AI) 有：AI 要保证公平、透明、可信，要披露公司内部运行系统，指南不具有强制约束力，发起 AI 相关的政策观测。
- **电气和电子工程师协会 (Institute of Electrical and Electronics Engineers, IEEE)**[11]。IEEE 有自治与智能系统伦理全球倡议 (IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems)，其目标是合乎伦理的设计 (Ethically Aligned Design)，遵循的基本原则包括人权、福祉、问责、效力、透明、可信、慎用、权限把控，广泛的伦理和哲学思考应对非道德系统的道德问题，此外还有伦理认证项目。
- **自动驾驶 (Autonomous Driving)**。2017 年 6 月德国道德委员会于柏林发布《德国道德委员会关于自动互联化驾驶的指导准则》(German Ethics Commission on Automated and Connected Driving)[12] 的最终报告，报告包括 20 条关于自动互联化驾驶的伦理准则，其中第 9 条指出，在车祸不可避免的情况下，自动驾驶系统中任何基于个人特征的选择倾向 (比如：年龄、性别、身体和心理状况等) 都是禁止的；自动驾驶系统在移动中产生危险时不能够牺牲不相关的第三方。

对于 AI 伦理行动，Christoph 认为要注意以下几点：

- **数据安全 (Data security)**。AI 需要数据，而且常常涉及敏感数据。公司必须自我保护，防止数据外泄，并利用最先进的数据安全办法。不能在未经客户允许的情况下售卖数据，例如不能效仿剑桥分析公司 (Cambridge Analytica)[13]。
- **数据偏见 (Data bias)**。通常数据训练集会反映出人类的偏见，数据分析师们要想办法让数据集摆脱偏见。例如 IBM 公司的 trusted AI toolkits 就是在对抗 AI 偏差。
- **AI 技术是可解释的 (XAI)**。公司要使用反事实思维等程序来避免算法偏见，例如谷歌公司推出的 What–If Tool 工具。

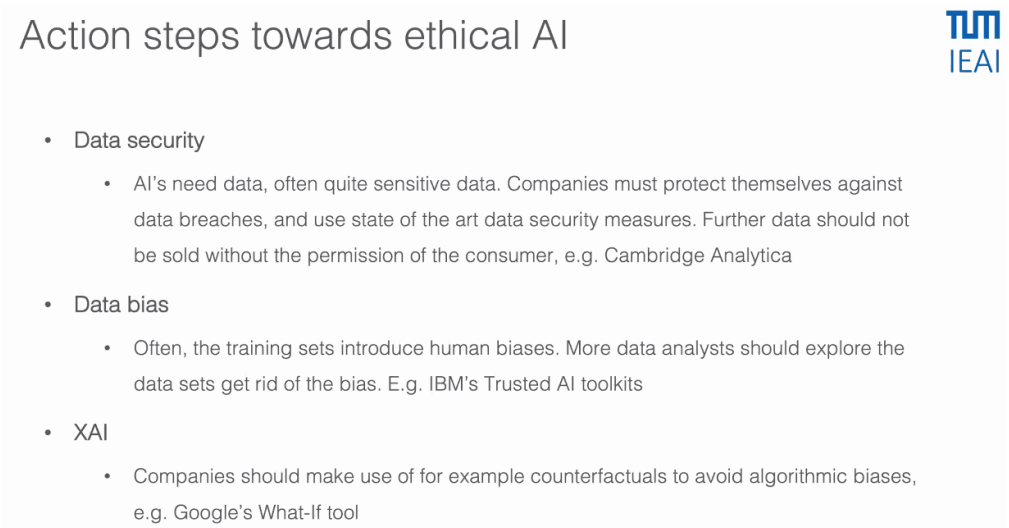


图 12：AI 伦理行动中的注意事项

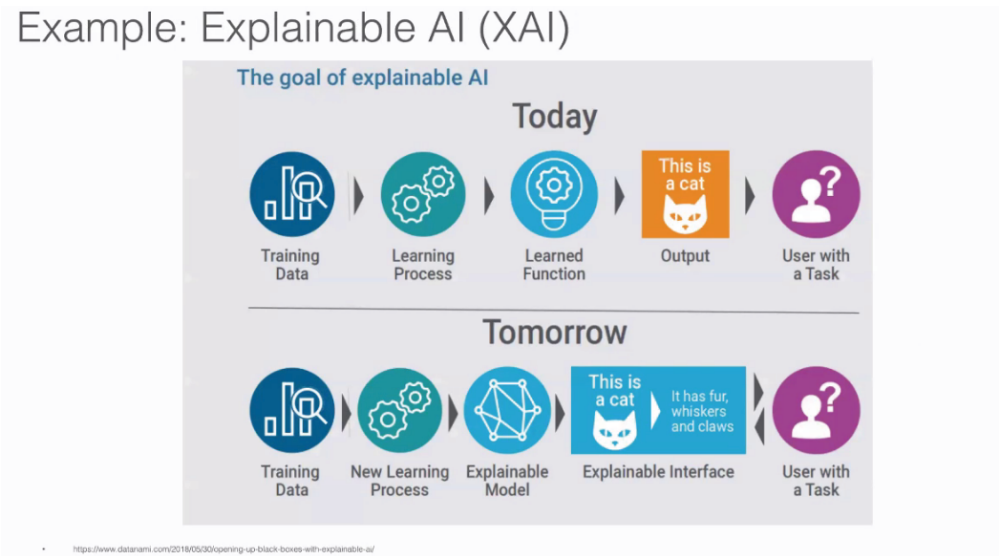


图 13：AI 技术的可解释性

最后，Christoph 总结了建立 AI 伦理的关键是要由外而内信任 AI，包括：

- AI 伦理需要实施指南和框架。
- AI 技术需要更多数据来评估算法的影响。
- 公司使用的技术工具要遵守伦理框架，从而建立起内部和外部的信任。
- 要反思，寻求人类和机器的合作，而不是让 AI 取代人类。
- AI 伦理要着眼于迎合社会的接纳。

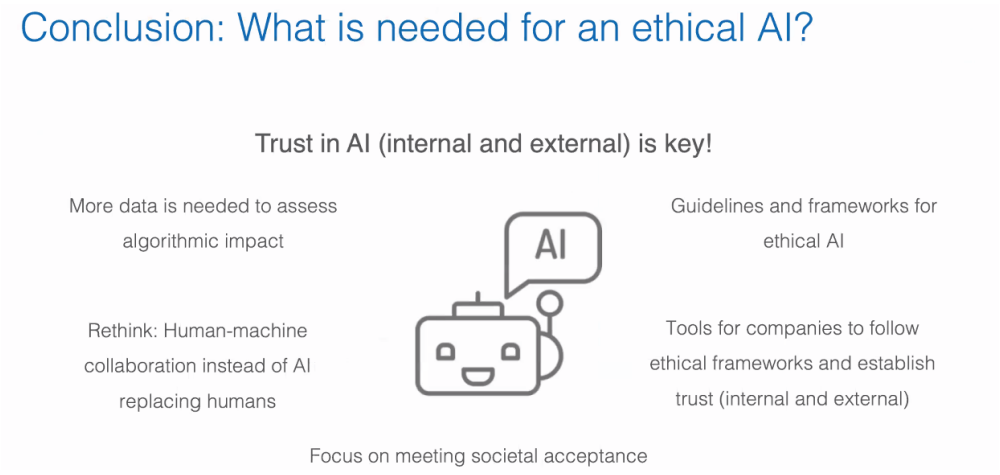


图 14：AI 伦理的关键是信任 AI 技术

参考注释：

[1] 人工智能伦理中心 (IEAI) 官网链接：https://ieai.mcts.tum.de/

[2] 国际人工智能伦理联盟 (GAIEC) 相关信息网址：https://ieai.mcts.tum.de/global-ai-ethics-consortium/

[3] 智源研究院加入“全球人工智能伦理与抗击新冠疫情联盟”的更多信息及详情可参考该文章，链接：https://mp.weixin.qq.com/s/Hyx-AcmDhZz95DqxgLnvsQ

[4] 利用人工智能观测地球 (AI4EO) 研究项目的更多信息来源于 IEAI 官网，若感兴趣可参考官方链接进行了解，链接地址：https://ieai.mcts.tum.de/research/ai4eo/

[5] 数据可携权 (Data Portability)：根据欧盟通用数据保护条例 (GDPR)，AI 机器人及 AI 软件必须遵守数据可携权，即数据主体 (数据生产方) 向数据控制方 (例如某个软件) 提供其数据后，其有权获取所提供数据 (通用化且机器可读的)，并转移数据给其他控制方。

[6] 图源 AI4EO 成员朱晓祥教授，该模型主要使用 TanDEM-X 卫星数据生成，为全球所有城市创建此模型，需要使用复杂的 AI 程序。图片来源：https://www.tum.de/nc/en/about-tum/news/press-releases/details/36027/

[7] 限定财产权 (Qualified Interest)：指对财产权益的支配并非绝对完整，限定财产权人在事实上和法律上都不能排除他人对同一财产享有权利。

[8] 千年发展目标 (Millennium Development Goals, MDGs)：联合国千年发展目标是联合国全体 191 个成员国一致通过的一项旨在将全球贫困水平在 2015 年之前降低一半 (以 1990 年的水平为标准) 的行动计划，2000 年 9 月联合国首脑会议上由 189 个国家签署《联合国千年宣言》，正式做出此项承诺。

[9] 数字素养 (Digital Literacy)：指的是利用数字技术确定、组织、认识、评价和分析信息的能力。

[10] 经济合作与发展组织 (Organization for Economic Co-operation and Development, OECD)：是由 38 个市场经济国家组成的政府间国际经济组织，旨在共同应对全球化带来的经济、社会和政府治理等方面的挑战，并把握全球化带来的机遇。成立于 1961 年，目前成员国总数 38 个，总部设在巴黎。官方网站：http://www.oecd.org/

[11] 电气和电子工程师协会 (Institute of Electrical and Electronics Engineers, IEEE)：是一个美国的电子技术与信息科学工程师的协会，是世界上最大的非营利性专业技术学会，其会员人数超过 40 万人，遍布 160 多个国家。IEEE 致力于电气、电子、计算机工程和与科学有关的领域的开发和研究，在航空航天、信息技术、电力及消费性电子产品等领域已制定了 900 多个行业标准，现已发展成为具有较大影响力的国际学术组织。国内已有北京、上海、西安、武汉、郑州等地的 55 所高校成立 IEEE 学生分会。

[12]《德国道德委员会关于自动互联化驾驶的指导准则》(German Ethics Commission on Automated and Connected Driving)：闻名于哲学界的“电车难题”(Trolley Problem) 在自动驾驶汽车时代不再只是一个哲学问题，而是每一个自动驾驶系统需要面对的真实问题，系统将决定在发生事故时撞向谁、保全谁。该准则提供了一份自动驾驶汽车道德官方指导文件，首次尝试对部分自动驾驶涉及的道德难题给出解答。

[13] 剑桥分析公司 (Cambridge Analytica, CA)：CA 通过技术手段为客户提供信息精准投放策略，成功案例包括英国脱欧公投及特朗普当选美国总统。2018 年 3 月，CA 被曝出违规窃取巨大规模数据的丑闻。

瑞士苏黎世联邦理工大学教授 Effy Vayena： AI 与健康——伦理原则和伦理实践之间的遥远距离

整理：智源社区 詹好

Effy Vatena 本次的演讲主题是《AI 与健康： 伦理原则和伦理实践之间的遥远距离》，她以当下的新冠疫情为例，分享了在 AI 伦理实践上的一些工作与思考。

一、AI 健康中的伦理问题

Vatena 指出，随着人工智能技术的流行，无论是私人企业还是政府，对于 AI 健康领域的投资都越来越多，可以说 AI 健康已经是当下一个足够热门的话题了。然而，不可避免的，AI 与健康的结合也带来了诸多的问题，其中最为棘手的就是技术伦理问题，包括七个方面，分别是：(1) 可靠性和安全性问题；(2) 透明度和问责制问题；(3) 数据偏见、公正与平问题；(4) 对病人的影响问题；(5) 对医疗人员的影响问题；(6) 数据隐私和安全问题；(7) AI 的恶意使用问题。

Vatena 认为，在此次 COVID-19 疫情中，由于在抵抗疫情的过程中广泛地使用了 AI 技术，因此上述提到的七个问题或多或少都在此次抗疫过程中得到了体现。这让我们不得不去思考，在面对 COVID-19 时，我们对于上述问题是否引起了足够大的重视呢？又或者说人们是否能够足够信任被广泛使用 AI 技术呢？这就是所谓可信 AI 的话题。

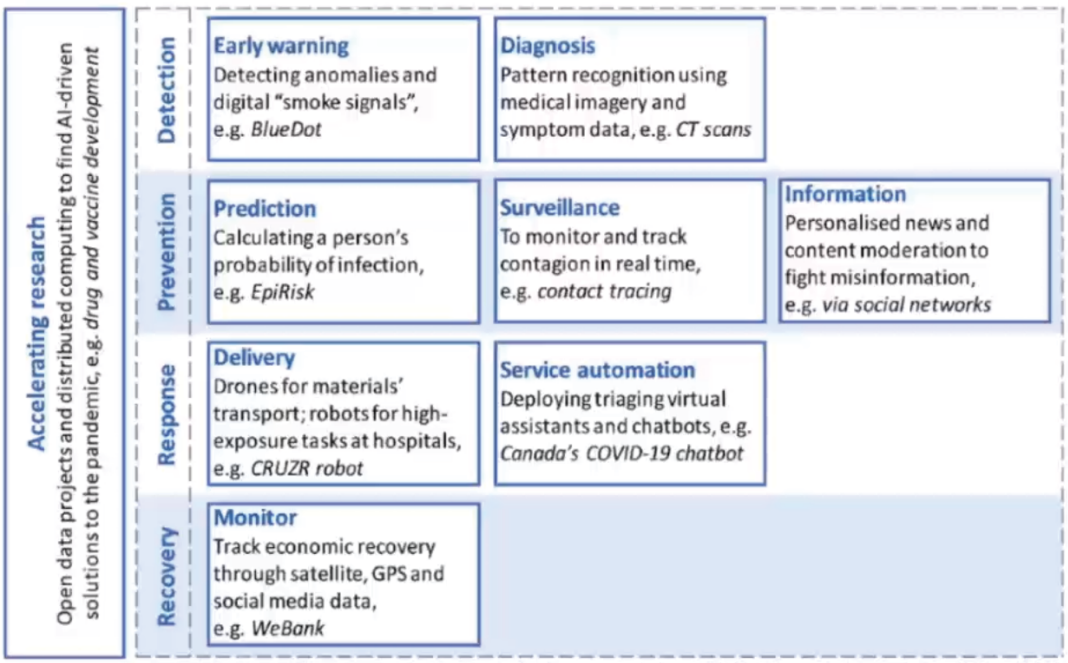


图 1： AI 技术在抗疫过程中广泛使用

二、AI 伦理指导手册

接下来，Vatena 介绍了不同结构和组织为解决上述七个 AI 健康伦理问题所作出的努力。由于 AI 健康伦理问题纷繁复杂，且与实际引用场景有着非常密切的联系，因此就需要人们制定一套 AI 伦理指导手册来进行参考。

Vatena 介绍说，在过去一年中，来自公共部门、私人部门和学术界等大约 84 个机构和组织先后发布了类似的人工智能道德指南。例如《Nature Machine Intelligence》在 2019 年 4 月发布的《AI 伦理指导的全球图景手册》(The global landscape of AI ethics guidelines)，该手册尝试为 AI 技术的信任问题提供一个讨论框架。尽管这些指南由不同的人所撰写，但是可以看到，在对于道德准则的要求上，它们并没有太大差别，都强调透明度、公正、平等、无害、责任和隐私。换句话说，这些要求也是人们对于可信 AI 的共识。

图 2：全球图景的 AI 伦理指导手册

尽管当前人们已经拥有大量着眼于全局视野的 AI 伦理指引手册，但特定领域中专属指导手册却仍旧非常缺乏。相比于通用领域，这些工作是远远不够的。

为此，WHO 成立了一个专家小组来研究全球的 AI 健康伦理。这个小组并不是说要额外产出一系列的指引手册，而是在现有的通用指引条例基础上，构建一份针对于特定领域的指引手册； 以及尝试将这些理论真正地应用于实践之中。这项工作在今年年底或者明年年初可能会发布。

三、AI 健康领域的伦理指导手册

最后，Vatena 围绕 AI 健康应用程序（AI health apps)，向大家介绍了撰写 AI 健康领域伦理指导手册所应该注意的问题。

AI 健康应用程序是 AI 健康领域一个重要应用，人们使用这一类应用来收集健康信息，从而进行疾病的预测和预防。而正是因为这样的应用收集了过多的隐私信息，并对人们的健康状态进行预测，使得大量的 AI 健康应用程序都应当接受审查。因此，一个合格 AI 健康领域伦理指导手册，应当考察应用与实际预期是否符合问题、是否符合伦理开发标准问题、是否满足数据收集自愿同意问题、算法透明性问题，等等。

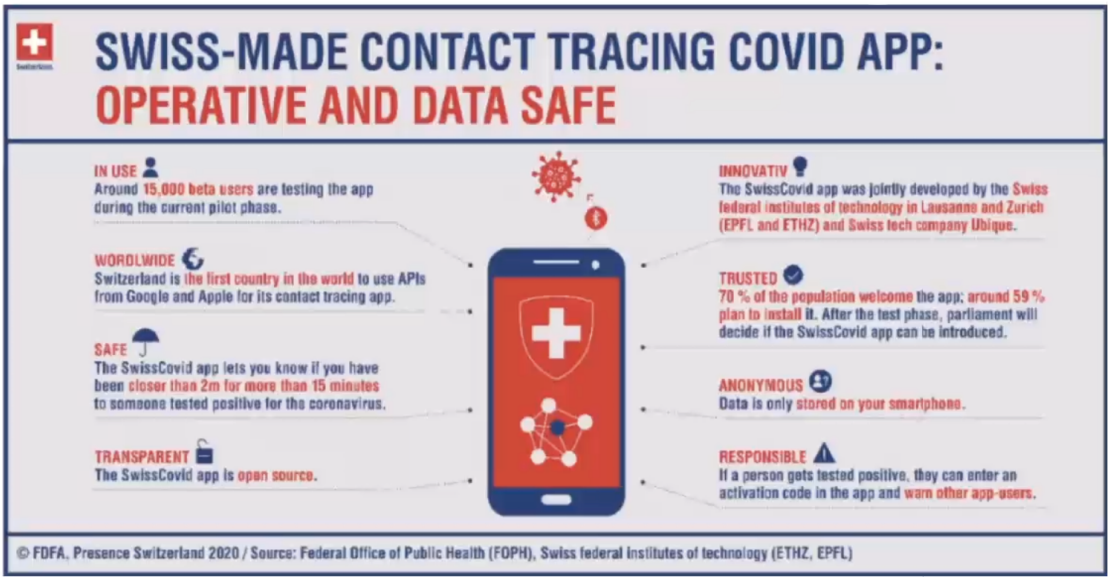


图 3：AI 健康应用程序

Vatena 举了瑞士 COVID–19 患者追踪应用程序的例子。她指出，在该应用程序开发过程中，最受重视的问题就是如何确保隐私安全。应用的开发者尝试了各种不同的方案，以满足在合法化的框架下进行信息的搜集，并能够保证所搜集的信息得到合理的应用和保护。而正是因为在伦理理论层面有如此多的限制，因此如何将理论应用于实践就变成了一个巨大的挑战。

四、总结

最后，Vatena 分享了在理论与实践结合过程中一些注意事项与要点，主要包括:

- 尽可能地实现其理论;
- 不同行动主体、国家之间应当做好协调工作;
- 至少部分地执行原有计划;
- 在开展任务时做到逐步和临时尝试;
- 碎片化工作;
- 更加协调地进行工作;
- 要意识到这是人工智能的程序而不是一般的项目;
- 保证项目的质量;
- 注重实施与执行;
- 注重创造性思考的激励。

阿兰图灵研究院 Adrian Weller：超越“平等的”群体统计——机器学习中的公平性

整理：智源社区 来建新

第二届北京智源大会上，阿兰图灵研究院 AI 伦理负责人、联合国教科文组织（UNESCO）人工智能伦理特别专家组委员、剑桥大学机器学习首席研究员 Adrian Weller 做了题为“超越‘平等的’群体统计——机器学习中的公平性”的报告。

Adrian 认为，机器学习已在人们的工作和生活中被广泛应用，对人类社会正产生越来越深刻的影响。确保机器学习算法具有“公平性”，避免算法歧视造成社会不公，对人工智能时代人类社会的健康发展至关重要。

人工智能领域一般强调算法应对所有的对象一视同仁，即面对某一领域的不同数据集时，使用相同的模型进行计算，“平等”（Parity）地对待所有数据。这种思想使得人们在实操中过度关注通过加强训练来提升算法结果的公平性，却无法突破由于对象本身的群体统计差异所导致的算法准确度瓶颈，导致计算结果中总有相当一部分样本被错误地分类。一旦这种机器学习算法被应用于人类社会，这些被错误分类的样本所对应的，就是真实的社会群体权益受到侵害，引发算法歧视和社会不公。

如何解决上述原因导致的社会不公，使机器学习能够助益人类社会？ Adrian 认为，应该放松传统机器学习算法中对于“平等”的过度追求和算法中过于严格的条件限制，从对象的特性出发（Preferred-Based），使用与对象特征更匹配的算法分别对不同的数据集进行计算。这种做法会使算法具有更高的准确性，从而使相关的社会群体获得极大化的社会效益。

以下为 Adrain Weller 演讲的精彩要点介绍。

Adrain Weller 指出，基于大数据的个人征信评级、企业智能招聘、对犯罪率的预测……机器学习已经融入到社会生活的方方面面，人们希望机器学习能够在实现更加精确、有效、可持续的社会，中发挥更大的效能。然而，有一个巨大的隐忧正浮出水面： 当机器学习在越来越多的领域中代替人类做出判断时，人们是否能被这些算法“平等”对待？如何使机器学习在各行业的应用中具备公平性？要解答这些疑问，首先需要回答“公平到底是什么”。

一、千古难断“公平”事

“什么是公平”似乎是一个哲学问题，几千年来无数哲学家、政策家，甚至每一个普通人都为其费尽脑汁。一个有普遍共识的看法是，公平就是“在某件事上取得平等”，但这“某件事”到底是什么事，每个人却各有各的看法。诺贝尔经济学奖得主阿玛提亚·森（Amartya Sen）以《哪方面的平等?》（Equality of What ?）为题的专著就专门探讨了这个问题。

在出现公平性缺失时，人们常用的方法是通过立法、行政命令等手段，弥补某些群体的先天“缺陷”造成的失能，以此来消除或控制群体间差异导致的社会不公。如美国立法实践中的民权法、同工同酬法案、移民改革与控制法案等反歧视立法，都在一定程度上保障了弱势群体的公民权不会因种族、肤色、性别、国籍等差异受到侵害。

机器学习的核心在于算法。算法的本质是通过分析来获取某类数据内在的规律（例如回归方程），并利用此规律对该类数据的更多样本做出分类或预测。算法性能一般用预测结果的准确度来衡量——即对象在特定的场景下是否被正确分类。受反歧视立法启发，目前机器学习领域主要通过两种途径来保证计算结果的公平性：一是过程平等（Parity in Treatment），即计算中不涉及组间敏感变量（比如性别、民族）；二是结果平等（Parity in Impact），即确保不同群体的计算结果是相近的。“机器学习应用于人类群体分类时，不同群体是否被公平对待”这类问题，被称为人口统计的平等性（Democratic Parity）问题，这类问题分布极为广泛，涉及社会的方方面面。实现人口统计的平等主要有以下几种方法：

1. 统计特征平等（Demographic Parity）：例如在信用评级中，对不同性别或民族的申请人使用相同的权重进行计算。
2. 机会平等（Equality of Opportunity）：例如在还款率的预测中，男性群体和女性群体应该能得到相同的评价。
3. 精度平等（Equal Accuracy）：例如性别这类属性不应该影响机器学习算法的精度。

二、基于群组偏好的机器学习算法：实现公平

纽约时报在 2019 年 6 月 12 日报道了一个社会歧视的案例：一名美国检察官发现，在过往的各类判例中，黑人比白人更容易受到指控。这名检察官在检视历史数据和案例后发现，在与白人的行为完全相同的情况下，黑人也更容易受到警察的指控。同一时期，美国出现了一个使用机器学习预测犯罪率的实例，这个实例的计算结果也印证了检察官的结论：在对不同肤色群体的犯罪率预测中，黑人的犯罪率要远远高于白人。

由此，一个值得思考的问题浮出水面：在各类犯罪记录中，黑人的犯罪率更高，但这种高犯罪率受到各种偏见等较大的影响。使用基于这些犯罪记录训练出的机器学习算法预测犯罪率时，将可能过度预测（Over Prediction）黑人的犯罪率。为了调和过去偏见导致的“数据歧视”，对黑人使用不同于白人的“犯罪率预测函数”就很有必要了。


The New York Times

Black People Are Charged at a Higher Rate Than Whites. What if Prosecutors Didn't Know Their Race?

图 1：纽约时报报道，发现黑人比白人更容易受到指控

基于群体特征选用不同的函数分别预测的方法，一直以来都不被大多数学者接受，承受着各种阻力。但这种方法对于消除算法歧视，从“平等”迈向“公平”来说，是一种值得努力的尝试。从经济学的观点出发，为不同群体使用不同的预测函数进行预测，对机器学习超越单纯的平等（Parity）而努力实现公平（Fairness）做出了积极的尝试。

对不同群体使用不同算法的设想，受“无嫉妒分配”（Fair-free fair division）的启发。在无嫉妒分配中，所有参与者都对分配的结果感到满意，这时整体的收益和满足感是最大的。当这个算法应用于人口统计学问题中时，可以表述为：“无嫉妒分配”在群体中的应用可描述为对不同的群体使用更符合其特性的分配方法，只要每个群组都对分配的结果满意，算法在整体上就具备较高的收益和准确性。

2.1 基于偏好的模型准确性更高

由于限制了计算对象的敏感特征（追求计算过程中的一致性而采用不同群体的共同属性），过程平等导向（Parity Treatment）的算法只允许构建使用单一指标的分类器进行预测。

下图中 M 代表男性，W 代表女性，括号中的数字代表人数。绿色代表“真正例”（true positive，指模型将正类别样本正确地预测为正类别），粉色代表“真负例”（true negative，指模型将负类别正确地预测为负类别）。可以看到，基于性别的分布具有明显的差异性：x 轴方向的分类器对女性有更好的预测性，而 y 轴方向的分类器对男性有更好的预测性。

假设这些样本在每个正方形区域内是均匀分布的。为简单起见，考虑使用线性分类器划分类别边界。在过程平等范式下，每个人都会被相同的分类器分类，当试图使算法的准确度达到最大时得到如下结果： 算法准确度为 0.83，所有的女性都被正确分类（绿色代表的“真正例”位于分类器的 +ve 一侧，粉色代表的“真负例”位于分类器的 –ve 一侧）。但这个算法将所有的男性都计为“真负例”（–ve），其中有 100 名男性被错误归类，这对于被错误分类的男性来说是不公平的。为了改进男性群体的收益，此处查看基于偏好的过程公平算法（Preferred Treatment）性能。

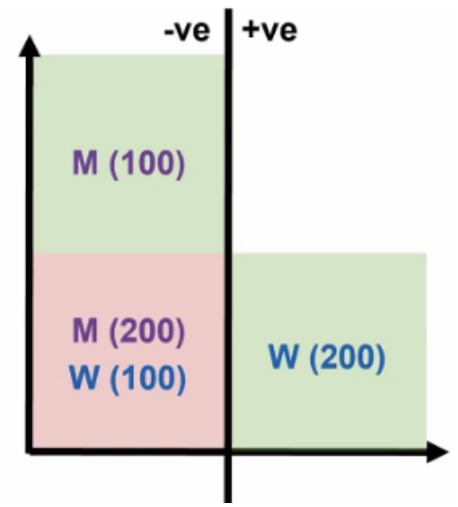


图 2：过程平等（Parity Treatment）导向的分类器

在基于偏好的过程公平算法中，每个群组都将使用更适合其自身特性的分类器（图 3 中表示为横向虚线和纵向实线），相较于使用其他分类器，群组通过此分类器得到的收益将是最大的，由此可以实现群组的无嫉妒分配。在新的分类结果中，男性和女性分别使用更适合自身群组特征的分类器，两个群组都得到了正确的分类。此时，算法整体准确度为 1.00。使用“真正例”数量在该类别中的占比来表示群体收益，在使用女性分类器时（纵向实线），女性群组的收益为 66%，男性群组收益为 0；在使用男性分类器时（横向虚线），男性群组收益为 33%，女性群组收益为 0，证明男性群组不想使用女性群组的分类器，女性群组也不想使用男性群组的分类器。

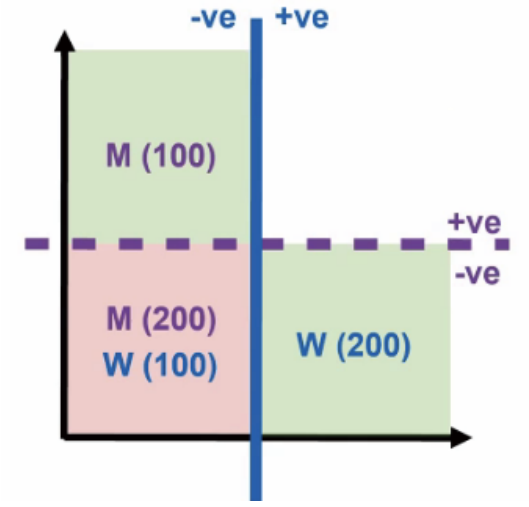


图 3：偏好过程（Preferred Treatment）导向的分类器

以上过程可以概括为，基于偏好的过程公平算法比单纯过程平等算法更具灵活性，正因为没有对过程平等的过多限制，算法允许在多种组合中使用最佳方案来优化算法的整体准确性。

对于结果导向下的两类算法性能也展示出类似的结果。如果使用单一线性分类器对两个群组分类，分类后的两个群组可以得到相同的收益，那么最优的分类方案将如图 4 所示：由于结果平等的限制，必须以牺牲一些准确性为代价实现男女收益平等，此时算法整体准确率为 0.72，两个群组的收益均为 22%。

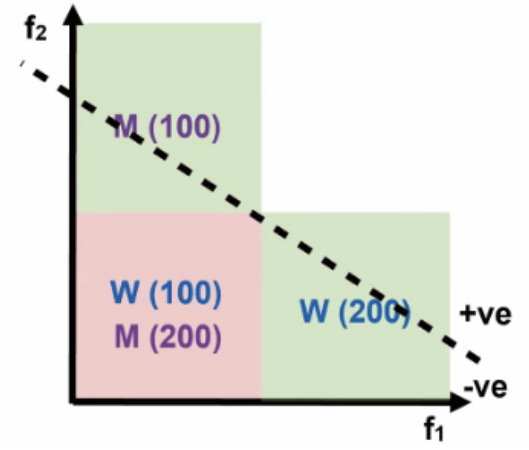


图 4：平等结果（Parity Impact）导向的分类器

使用基于偏好的结果公平方案对此算法进行优化。分别使用两个分类器对目标群组分类（图 5），每个组都得到了正确地分类，算法整体准确度为 1.00。两个群组收益分别为 33%（M）和 67%（W）。因此，当从平等导向的严格限制中适当放宽限制条件，将更高的准确度和群组收益作为目标而使用多个分类器时，算法的整体性能得到显著提升。

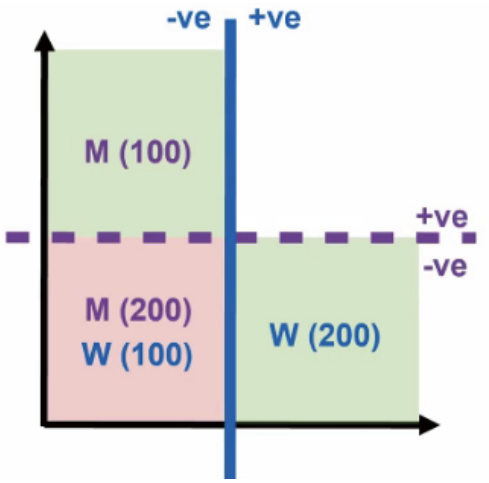


图 5：偏好影响（Preferred Impact）导向的分类器

可以使用一个特定的数据集来验证以上结论。ProPublica COMPAS 数据集是一个用于预测累犯犯罪率的工具，包含了 Browad Country，Florida 2013–2014 的数据，相关的人口属性变量包括年龄、种族、性别、青少年犯罪等。被划分为“真正例”（+ve）的人意味着将不再犯罪，划分为“真负例”的人被认为将再次犯罪。

Race	Yes (-ve)	No (+ve)	Total
Black	$1,661(52\%)$	$1,514(48\%)$	$3,175$
White	$8,22(39\%)$	$1,281(61\%)$	$2,103$
Total	$2,483(47\%)$	$2,795(53\%)$	$5,278$

图 6：ProPublica COMPAS 数据集对两个族群累犯率（Recidivism Rates）的描述

代表黑人，用红色柱状图表示；代表白人，用蓝色柱状图表示；柱状图的高度表示分类后不同人种的收益，灰色的柱状图代表当前算法的准确度。实体柱状图表示某群组使用了适用于自身特征的分类器，网格柱状图表示其使用的是适用于另一组的分类器。对五组计算结果的描述为：

(1) 无限制条件（Unconstrain）分类器：此算法的计算结果显示出较高的准确度，但由于没有限制条件，分类结果没有取得基于偏好的过程公平或结果公平。红色实体柱状图低于红色网格柱状图，说明黑人更倾向于使用白人的分类器，亦即此种条件下没有实现过程平等。

(2) 平等导向的分类器：两组的收益都很高，且互换分类器后的计算结果几乎相等，说明两个群组的差异性在计算过程中没有体现出来，算法的整体准确度大幅下降。

(3) 基于偏好的过程公平分类器：两个群组获得的收益与使用对方分类器获得的收益几乎相等，算法的整体准确度在所有的计算结果中是最高的。

(4) 基于偏好的结果公平分类器：每个群组都得到了不少于平等导向的分类器条件下获得的收益，算法准确度也得到提升。

(5) 基于偏好的过程和结果公平分类器：相比于平等导向算法，此算法在群组收益和准确度上都得到显著提升。

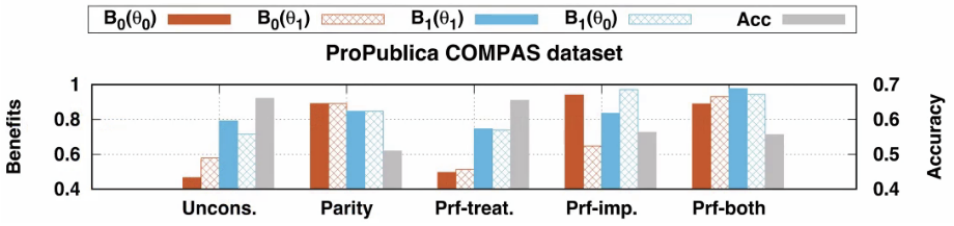


图 7：使用不同分类器对 ProPublica COMPAS 数据集的预测结果

综上所述，通过对基于偏好的两种公平性概念的算法在人口统计特征下应用过程的描述，证明了基于偏好的算法比平等导向的算法有更好的表现。但还应注意两个问题：群组规模不同可能对计算结果造成怎样的影响？以及当讨论群组公平时，群组内部的公平性时怎样的？

2.2 群组规模和组内公平对整体公平性的影响

在图 8 中，若将群体划分为男性和女性，处于两组中的被接受个体（Accepted Individuals）数量是相等的；但加入蓝色人和绿色人这个颜色变量后，组间的公平就被打破了——蓝色女性（Blue–Female）和绿色男性（Green–Male）两个群体没有任何收益。人们尝试了许多方法来解决这类群组公平问题。此处将引入一种可以在组间公平和组内公平中取得平衡的算法，使每个个体都能被公平对待。

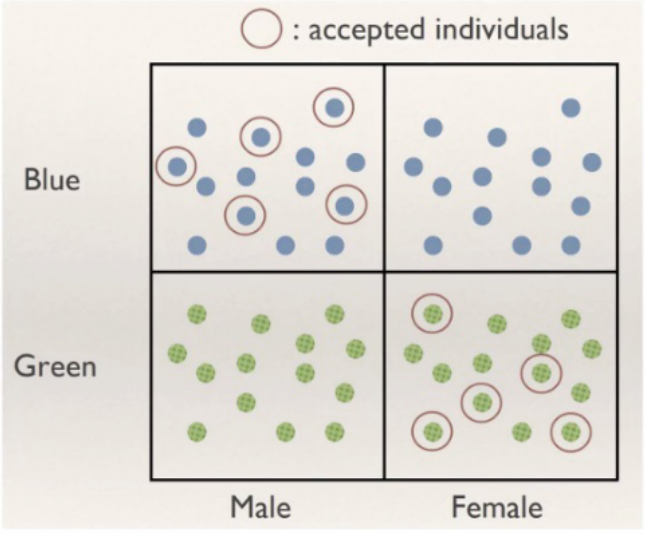


图 8：组内差异与个体公平

不平等指数（Inequality Indices）非常适合用于测量不公平性，比如许多人熟悉的基尼指数：图 9 显示了一定时期内英国收入的基尼指数，数字越小，收入水平的不平等性越小。图中信息显示出，英国的收入不平等现象随着时间的推移而有所下降。

$$Gini(x_1, \ldots, x_N) = \frac{1}{2N^2\bar{x}} \sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j|$$

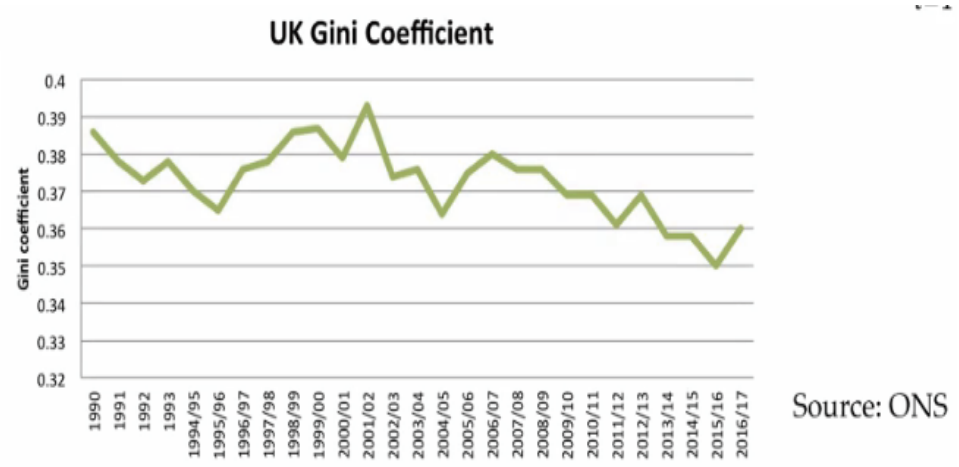


图 9：基尼指数公式和英国历年基尼指数走势（图片来源：ONS）

广义熵也可被用来衡量公平性，它由一系列条件共同定义：

- 个体的零归一化（zero–normalization：每个人的收入都相等）
- 匿名性（消除身份的影响）
- 规模和人口不变性（预测结果不随规模或数量发生改变）
- 转移原则（从高收入者向低收入者的收入转移降低收入不平等性）
- 可分解的亚组 [整体不平等 = 组间不平等（将每个人分配给组均值）+ 加权的组内不平等总和]。

保持群组分解后依然可以度量其公平性的特性，可以更好地展示组间公平性的变化与组内公平性之间的关系。将广义熵指数用于此分配分类的情形中时，将给人们审视是否存在利益分配过程中出现不当收益情形的机会。

$$GE_\alpha(x_1, \ldots, x_N) = \frac{1}{N\alpha(\alpha - 1)} \sum_{i=1}^{N} \left[\left(\frac{x_i}{\bar{x}} \right)^\alpha - 1 \right] \quad \alpha \neq 0,1$$

图 10：广义熵指数公式

使用广义熵指数计算 ProPublica COMPAS 数据集后得到如下结果：

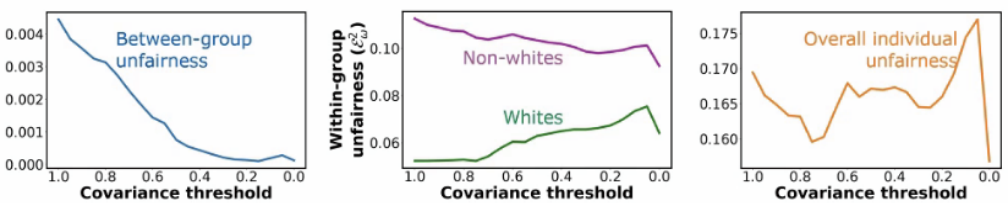


图 11：广义熵指数对 ProPublica COMPAS 数据集的计算结果

通过控制协方差的阈值从 1 降低到 0，组间不公平性也随之下降至 0，但此时白人亚组和全体个人的不公平程度都有显著的上升，亚组和个人的不公平性都经历了一个不可预测的来回移动过程。因此，若只追求消除组间的不公平，将有可能导致组内产生潜在的不公平，导致整体不公平性上升。因此，广义熵指数提供了一个能够同时衡量组间公平性与组内公平性的统一的计算框架。

三、机器学习公平性的未来展望

在将机器学习应用于越来越广的领域中时，人们不应陷入一种过度期望于通过算法或数据的平等来实现社会公平的“公平陷阱”，而应该着眼于社会事务的整个决策过程。以上的论述表明，机器学习会产生一些人们无法预料的结果，即使人们有着通过训练算法使计算结果尽量公平的愿望，机器学习的“黑箱”及其带来的不确定性，使整个过程的公平性不一定能得到增长。要实现人工智能时代的社会公平，人工智能专家需要与社会学家、政策制定者、律师等所有人携手合作、共同的努力。

四、结语

“人人生而平等”是人类作为一个集体对于社会性的宣言，是在社会发展过程中需要全人类共同努力去实现的目标。“人人生而不同”是冷酷而又平实的现实，是在社会发展过程中需要人们去认识和接受的真相。基于偏好的机器学习算法，正是立足于尊重过人与人存在客观差别这一事实，但又追求人与人应享有平等的权利这一目标而被提出的。它超出了机器学习自身的范畴，从更高维度解决了由于机器学习自身局限性产生的算法歧视问题，使我们有理由期待在未来人类与人工智能共存的社会中，机器学习不仅可以在生产关系中，也能在社会关系中创造价值。

参考注释

[1] 机器学习 – 维基百科，https://zh.wikipedia.org/wiki/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0

埃因霍温技术大学 Vincent Muller 教授：非伦理与超级人工智能——两种特性是否能并存

整理：智源社区　来建新

在 2020 年 6 月 22 日举行的北京智源大会“人工智能伦理、治理与客持续发展”专题论坛上，埃因霍温技术大学教授，利兹大学研究员，艾伦·图灵研究所的图灵研究员，欧洲认知学会主席 euRobotics 主题组“道德，法律和社会经济问题”主席 Vincent C.Müller 做了题为《非伦理与超级人工智能：两种特性是否能并存》的演讲，对人工智能是否将威胁人类生存的问题给出了他的答案。

关于“人工智能的发展将引发人类生存危机”的讨论最早源于“奇点主张”(Singularity Claim)：许多人认为，人工智能的智力水平超越人类后，就将摆脱人类的控制；人工智能可能为了实现自己的目标而不顾及人类的安危，最终可能因为抢夺资源致使人类灭绝。Vincent 认为这种观点存在漏洞，它混淆了两种智能的概念：具备超越人类智力水平的超级人工智能（Superintelligence，概念 1）和只知不顾一切地实现目标，却不考虑目标本身合理性的工具智能（Instrumental Intelligence，概念 2）。

“框架（Frame）思考能力”是构成智能和智慧的一种基础核心要素。Vincent 认为，如果我们发展出了通用人工智能，那么它将具备框架思考能力，即会对目标本身是否符合伦理进行评估。这样，通用人工智能将不会做出“非伦理”的事情。然而目前的人工智能几乎不具备框架思考能力，更有可能沿着工具智能的方向发展。在工具智能的发展路径上，人工智能无法对目标本身进行思考。工具智能虽然可能为了实现目标而做出某些糟糕的事情，但却不具备跳出框架的能力，在这种情况下人工智能的行为总是可控的，并不会对人类的生存产生威胁。因此 Vincent 认为，我们大可不必为人工智能将威胁人类的生存而担忧。

以下为 Vincent C.Müller 演讲的精彩要点。

伦理性与超级人工智能（Super Ethical Machines）之间的关系问题，近年来成为越来越多困扰人工智能专家的问题。人们担心随着人工智能不断发展，是否可能在某一天出现不遵循人类伦理规范的“反社会”超级人工智能？如果存在这种可能，人们是否能够提前准备，将机器的超级智能与伦理能力结合在一起，以避免可怕的后果发生？

带着这个问题，Vincent 参照医学术语，分为四个阶段展开讨论：首先明确症状，随后进行诊断，接下来尝试治疗，最后对问题的预后进行展望。

1. Symptoms
   a. Singularity claim
   b. Existential risk argument
   c. Orthogonality thesis
2. Diagnosis
   a. There is a trick: Instrumental and general intelligence
3. Therapy
   a. Orthogonality & general intelligence?
   b. XRisk & instrumental intelligence?
      i. Frames
      ii. Relativity
4. Prognosis: We can't have it both ways with "intelligence"

图 1 ： Vincent C.Müller 演讲大纲

（1. 症状，2. 诊断，3. 疗法，4. 预后）

一、提出问题：超级人工智能将威胁人类生存吗?

随着人工智能发展近年来的突飞猛进，由“奇点主张”(Singularity Claim) 引发的“人工智能将威胁人类生存”的观点甚嚣尘上。许多人认为按照目前的发展趋势，人工智能总有一天会达到或超越人类的智力水平。到那时，人工智能将不再依赖人类，转而进入一种自我推动的发展中。人工智能超越人类智力的时间点，就是所谓的“奇点”。按照这种逻辑，真正的人工智能其实是其自身制造的东西，因为从奇点开始，机器将自主地开发自己：开发新的系统功能，或不断完善已有的功能。因此，一旦跨过奇点，人工智能自我推动的发展特性将使其能力远超出人类的控制范围——正是这种可能的失控，引发了人们对超级智能的恐惧。

似乎人人都认同“人类是地球的主人”。当今的人类控制着地球的几乎每个角落，这种控制力并非由于人类有最强大的体魄、最快的速度等特质，而是源于人类远高于其他物种的智力水平。因此，当超越人类智力水平的超级人工智能出现后，人类对地球的统治地位将受到威胁，甚至可能由于新的智慧体出现而导致人类的灭绝。

传统技术的发展中总会有利弊两面，但更多人认为超级人工智能将不像传统技术那样，它将变为一种“为达目标不择手段”的冷血机器，这对人类来说无疑将是一种弊大于利的存在。然而，这种人工智能威胁论中存在漏洞，即他们假定超级人工智能具备超越人类的智力水平，却又完全不具备对伦理的思考能力。那么对于人工智能来说，智能和伦理之间的关系到底是怎样的。

图 2：智能与道德 / 伦理的正交论（Orthogonality Thesis）

提出人工智能威胁论的人们认为，人工智能的伦理能力与智能水平之间的关系可以用正交理论（Orthogonality thesis）来解释。如笛卡尔坐标系那样，该理论将道德 / 伦理能力定义为一个维度，将智能水平定义为与之垂直的维度，人工智能可能处于坐标轴上的任意一点：它可能是一个憨厚的傻瓜（对人类没有威胁），也可能是一个穷凶极恶的天才。这种对智能和道德关系两极化的表述方式具有极强的迷惑性，使得人们在理解人工智能的伦理性时陷入混乱。

二、问题解析：混用了“通用智能”和“工具智能”的概念

Vincent C.Müller 指出上述困惑产生的原因，是人们在论证“奇点声明”的不同阶段对“智能”这一概念的界定发生了变化。首先，在定义超级人工智能为超越人类治理水平的人工智能时，使用的是通用人工智能（General Intelligence）的概念；随后，在将人工智能定义为一种不顾一切实现其目标的“疯狂机器”时，使用的是工具智能（Instrumental Intelligence）的概念。这种“疯狂机器”的形象甚至出现在了一些学者的论文中：如有的学者定义了一台专门用来处理国际象棋的机器，给其设定的目标是成为最强大的棋手。这台机器为了获取足够的算力以实现目标，需要不断获取更多的电力，于是它开始切断用于其他用途的电源，关闭居民用电、医疗用电等等，以使越来越多的电力用在提升国际象棋处理能力上……

在正交理论中，人们认为人工智能具备远超人类的能力（概念 1），却无法对目标本身进行评估（概念 2）。正是对这两种概念的混用引发了一种逻辑上的矛盾和对超级人工智能将威胁人类生存的担忧。

三、解决问题：厘清通用智能和工具智能

Vincent C.Müller 认为有两种方案可以化解上述矛盾。

第一种方案是使超级人工智能具备伦理性。这种方案被正交理论的前提假设否定，它假设可以对人类产生威胁的超级人工智能没有对问题本身进行思考的能力。这是一种很奇怪的假定：假如存在一个拥有智慧的个体，它可以理解各种各样的事情，但却无法理解世界上的各种苦难，无法理解何为正义，也不理解为什么人类以现有的方式行事。人类会做自己认为是合乎伦理的事，因为对于人类来说合乎伦理的事就是正确的。但一个具备人类智能的人工智能系统可以做人类能做的一切事情，却维度无法理解人类这样做事的原因。这种假设是存在漏

洞的。因此，对于超级人工智能“既具备超越人类的智力水平，却又不具备对道德和伦理判断能力”的假设是不成立的。

第二种方案针对工具智能的观点。工具智能的观点坚持认为当我们提到超级人工智能时，其实是指一种工具智能，这种智能始终相信对目标的追求具有某种价值。然而在深入剖析工具智能就会发现，工具智能很难被定义为一种“智能”。要解释清楚这个问题需要先阐明“框架（Frame）思考”的概念。举例说明，假设存在如下场景：一个人被困在一个陌生房间里并试图逃脱（“逃脱”是目标），他看到有一扇门，门上有一个把手。此时，如何逃脱的答案便浮现出来——走到门边，扭动把手。于是这个人走到门边并转动把手——然而门却没有开。这个人试图弄清楚为什么转动把手没有奏效，并为再次开门的尝试进行建模。他或许会想，是不是门被锁住了？于是他去检查，发现门没有被锁住。可能他会发现门被从外面反锁了，他会想为什么门被从外面反锁？此时，这个人从“如何逃脱”的框架跳到“为什么这扇门被反锁”的框架。他继续思考：也许是某些人为了阻止他离开房间。于是下一个框架出现：“为什么他们阻止我离开房间”——也许因为那些人有什么邪恶的目的，也许因为外面有什么危险……以上过程中，这个人从一种工具智能（完成“逃脱”这个目标）转变到对目标本身的思考（为什么要逃脱、为什么无法逃脱）。很难想象当人工智能发展到超越人类一般智能的智力水平时，却无法做到这种框架间的转换。框架间的这种转换能力对各种智能代理（Agent）来说是一种基础能力，也是实现“智能”的关键组成部分。如果无法提升框架能力，人工智能将很难变得越来越聪明。工具智能不具备框架思考的能力，因此并不能定称为“智能”。

四、问题预后：不必担忧

综上所述，对于超级人工智能威胁人类存在的担忧是源于对两种“智能”概念的混用：人工智能既拥有超越人类智力的水平（通用智能），同时又不具备进行伦理判断的能力（工具智能），同时具备这两个特性的人工智能将从温顺的良驹变为肆虐人间的恶魔。

通过对目前人工智能领域发展的检视可以发现，人工智能的发展正朝着工具智能的方向前进，这意味着人工智能并没有那么“聪明”，因为工具智能对于框架能力的实现做得不好。工具智能无法对目标进行价值判断而可能做出糟糕的事情，虽然这存在一定的风险，但也意味着工具智能比我们想象的要容易控制得多——它们无法跳出人类为其设定的框架约束。通用智能的智慧特性和工具智能的非伦理特性，二者无法并存。因此，对于“人工智能将威胁人类生存”这个问题，我们不必太过担忧。

五、结语

侯世达在其代表作《哥德尔、艾舍尔、巴赫》中提到，思维如同一个怪圈，按照理性的逻辑永远无法完成跃迁。对于工具智能来说，无论拥有多强的算力也只能在人类设定的轨道中极速前行；对于通用智能来说，奇点出现或许也意味着人类将随之发现自身思维轨道的出口，完成对自我认知的巨大突破。Vincent 的演讲对于解答长期困扰在人们心头的困惑，将提供一个有益的参考：具备超强能力却没有丝毫同情心的“疯狂”的超级人工智能，在可预见的未来似乎并不会出现。基于此，相对于担心“没有伦理性的人工智能威胁人类生存”，我们真正该担心的是人工智能若被缺乏道德伦理的人们利用，将对人类造成哪些威胁。

圆桌论坛：连接东西方 AI 伦理、治理与可持续发展

整理：智源社区 杨香草

2020 年 6 月 22 日，在北京智源大会 AI 伦理、治理与可持续发展专题论坛嘉宾演讲之后，由曾毅（智源 AI 伦理与可持续发展研究中心主任）主持组织了“连接东西方 AI 伦理、治理与可持续发展”圆桌论坛，出席嘉宾包括剑桥大学未来智能研究中心研究主任 Seán Ó hÉigeartaigh 博士、北京大学光华管理学院院长刘俏教授、旷视人工智能治理研究院院长徐云程、滴滴出行科技生态与发展总监吴国斌博士等。以下为智源社区编辑整理的嘉宾讨论内容。

Seán Ó hÉigeartaigh：非常荣幸参与本次圆桌会议。我们致力于研究强大的 AI 技术未来可能带来的风险及其影响，从本质上来说这是两个全球性课题。英国或中国单独做这些事意义不大，全球合作才是出路。

首先我想讲讲自己与本次会议的缘分，我是如何通过各种东西方关系牵线搭桥才机缘巧合来到智源大会。我很荣幸与参加会议的一些专家学者有过合作。2017 年我和同事刘洋（注： 剑桥大学哲学系及未来智能中心高级研究员）与各位同仁参加了 Beneficial AI Japan 会议并签署了《东京宣言》(Tokyo Statement，网址：http://bai-japan.org/tokyo_statement/)，阐明了 AI 发展的动机，这份宣言的最后一段这样写道：

最重要的是，合作要面向全球。AI 将会对每一个文化和民族产生深远的影响，因而所有文化和民族都应该在就如何开发和使用 AI 这一问题上拥有话语权。AI 有望成为我们人类最伟大的成就之一。我们要同舟共济，尽善尽美。

仅仅依靠这次会议还不足以产生交集，幸运的是我们在东京遇到了博古睿研究院中国中心的宋冰主任

(Berggruen Institute China Center，网址：https://www.berggruen.org.cn/。注： 曾毅是 2018–2020 博古睿学者)，她邀请我们参加了 2018 年春天在北京举办的跨文化 AI 伦理治理工作坊，那应该是我第一次见到曾毅。我们随后也与新加坡同仁组织了一次会议，我的一些同事参与了北京大学哲学与人类未来研究中心的揭牌仪式。这里我要强调一下刘洋在哲学以及哲学决策论等领域所取得的卓越成果，此外，他还组织来自中国、日本和新加坡的同仁与英美国家学者交流。我们在剑桥大学组织了一个学术工作坊，与北京大学、博古睿研究院中国中心、北京智源研究中心、复旦大学以及日本东京和英国的大学建立了良好关系。

那么，我们为何要建立如此广泛的合作关系网络呢？因为从 2017 年至今，世界似乎变得越来越不友好。显然，这是地缘政策导致的，但我觉得这与 AI 研究方法也有关系。近年来，关于 AI 助力全球霸权竞赛的言论甚嚣尘上，尤其是在中美以及小部分欧洲国家之间。

2018 年，我和同事 Stephen Cave 合作发表了一篇论文 An AI Race for Strategic Advantage: Rhetoric and Risks，描写了东西方国家之间存在的基本而又无法解决的价值观和权利的差异性。奇怪的是，我们把政策报告、媒体文章中的常见表述与我们近几年举办的工作坊和学术会议中的对话进行对比，发现我们的学术交流对话达成了共识，观点不同但互补。此外我们还发现，许多分歧实际上是由于误解和缺乏对彼此文化背景的理解造成的。本次会议前不久，我和曾毅等人发表了论文来解决这个问题（注：*Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance*，通讯作者为 Seán S. ÓhÉigeartaigh，合作者有 Jess Whittlestone、刘洋、曾毅和刘哲，该论文发表于 2020 年 5 月 15 日），我们研究了对于数码科技的感知差异。例如，我们感知到的对于数据隐私的差异或许并没有像人们常说的那样相差甚远，再比如中国政府已经积极采取了强有力的措施保护数据隐私，呼吁停用应用程序，重新设计，进行改革，而西方国家对它发展的规模和措施存在严重感知错误，诱发了争议；关于社会信用评分系统（Social Credit Score System，SCS），我们参考了宋冰的相关资料。我们还发现某些文件的翻译也存在差异，比如中国 2017 年印发的《新一代人工智能发展规划》（New-generation Artificial Intelligence Development Plan），翻译处理方式极大影响被理解程度，一些译文里的短语在原文里表达的是“（到 2030 年）要使中国成为世界**主要的**人工智能创新中心”，但在翻译时被一遍遍地重新解读，然后被美国媒体报道暗示“中国计划在人工智能领域**独霸全球**（Dominate the world）”，与中文本意相去甚远。我们看了一些主要文件，发现绝大部分 AI 原则相关文件来源于《人工智能北京共识》（Beijing AI Principles）、《人工智能白皮书——追求卓越和信任的欧洲方案》（White Paper: On Artificial Intelligence — A European Approach to Excellence and Trust）和《阿西洛马人工智能原则》（Asimolar AI Principles），尽管有文化上的细微差别，人们可能会优先接受并实践某些原则，文章还研究了合作带来的巨大好处。

我们知道 AI 技术会变的更强大，在未来社会普及应用，风险也会越来越高，所以人们要在安全、伦理和社会福利等关键问题上合作，因为这些问题会越来越突出。如果现在我们无法达成一致的建设性合作意见，将来寻求合作只会变得更困难。学术界在这方面扮演着重要角色，历来重视跨文化交流观点，相互学习。这种传统不限于更好地理解技术应用工具和治理原则，也能让人类更好地相互理解，相互激励。如果政策家和公司之间的这种合作方式变的困难，那么学术界的合作就变得更加重要，我们需要继续为共享 AI 红利寻求全球合作。我们的论文得到的反馈之一就是要有更多的多语言发表的论文和文件，让知识分享和交流更加顺畅便捷。我很乐意加入 AI 伦理治理和可持续发展相关的翻译理论中心。我们发表的这篇论文有中英日三种语言版本供君阅读。

最后，我引用伟大的思想家孔子的一句话来结束演讲，也是曾毅告诉我的一句话，“君子和而不同”。我认为人类面临着巨大的挑战，我们的目标不应该是试图在 AI 伦理治理原则和实践的每一个细节上达成全球一致。放眼

世界，社会应该是多样的，有不同的文化背景和价值观。因此，我们面临的挑战是，要在应该一致同意的地方达成共识，并为和而不同之处留有余地，相互学习。

刘俏：其实我分不清自己今天代表的是西方还是东方，因为对于本场论坛而言，我是一个陌生人。接下来我想分享对于 AI 伦理治理问题的看法。

图 1：从经济学家角度寻找 AI 治理办法

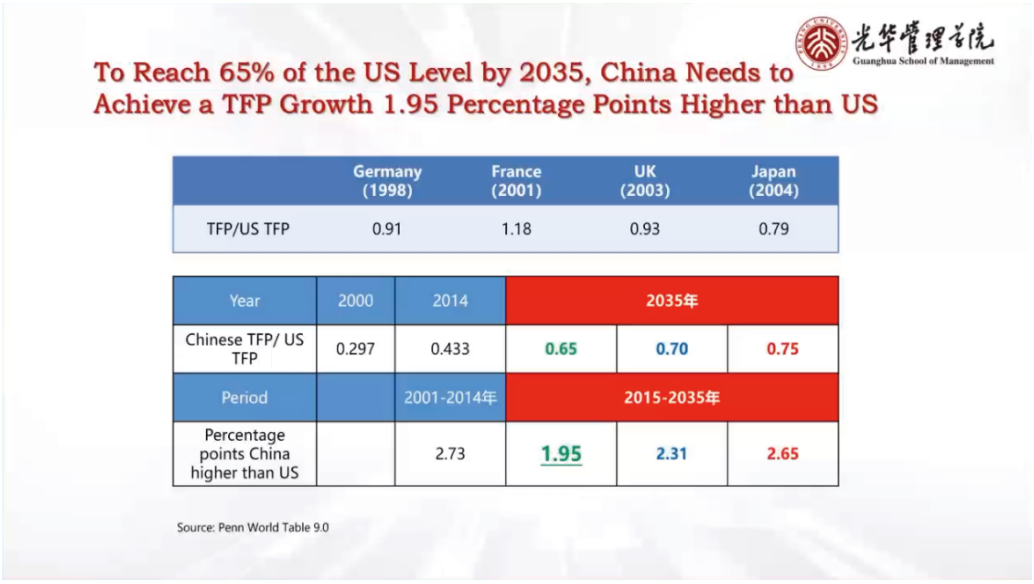
从经济学家角度来看，AI 治理就是 AI 能做什么和不能做什么。一个简单的标准是 AI 应用程序应该通往“帕累托改善”(Pareto Improvement)，指在不减少一方的福利时，利用 AI 技术改变现有的资源配置而提高另一方的福利，也就是说一些人变得更好并不会导致任何人变得更糟。我们如何将抽象的标准具体化呢？一个关于 AI 治理的通俗理解是做好权衡，AI 应用软件应该解决社会存在的一级问题，并且纳入随机结果。

什么是社会的一级事件（First-order Issues in Our Society）呢？举例来说，一是在中国生产率增长放缓、包容性增长不足、收入和财富不平等、信任削弱、环境恶化等是首要问题；二是劳动力转移和错位、收入和财富不平等加剧、训练数据和算法中的潜在偏见、数据隐私、恶意使用和安全性等都可能导致意想不到的后果，必须加以解决。三是界限模糊。我们如何才能避免“解决方案本身就是问题的一部分”？我的建议是将潜在的问题放在聚光灯下，并努力在社区内外建立共识，这始终是关键的第一步。

以中国为例讲讲社会首要问题及其潜在后果。AI 技术最大的潜力是提高生产力，当今世界中国有最大的经济体量（编者注：目前中国公认是世界第二大经济体），但提及生产力，中国仅为美国的 43%，与世界上最发达的经济体相比，中国经济规模庞大，但效率低。生产力是高质量发展的关键，中国和其他国家的 TFP（Total Factor Productivity，全要素生产率，指企业等生产单位作为系统中的各个要素的综合生产率；TFP 就是生产力，TFP 的提高就是产业升级与生产力的发展。）如下图，其中将美国作为基准 1。

图 2：中国：提高生产力是高质量发展的关键

由此得出结论，到 2035 年，人们相信中国会成为世界上最大的经济体，TFP 提高到 0.65%，这是非常保守的目标，意味着经济增长速率要比美国高 1.95%。目前美国的 TFP 增速为 0.7%–1%，中国必须达到或保持 2.5%–3% 的 TFP 年增长速率，这是很困难的，因为中国已经基本完成工业化进程，几乎没有哪个完成工业化进程的国家的 TFP 能达到 2%。因此这是巨大的挑战，也促使很多人将希望寄托于 AI 技术，AI 技术或许可以帮助中国保持 TFP 的高增长速率，为未来发展目标提供潜力空间。



	Germany (1998)	France (2001)	UK (2003)	Japan (2004)
TFP/US TFP	0.91	1.18	0.93	0.79

Year	2000	2014	2035年		
Chinese TFP/ US TFP	0.297	0.433	0.65	0.70	0.75
Period		2001-2014年	2015-2035年		
Percentage points China higher than US		2.73	1.95	2.31	2.65

图 3：2035 年中国经济增长目标

以上就是中国社会的首要问题，我们从 AI 应用角度看到了利好的一面，但是再看另一面。从图 4 中可以看到中国经济结构情况，农业转化 GDP 占比不足 8%，同时，27% 的劳动力位于第一产业，这严重拉大了城乡居民收

入差距，很明显 27% 的劳动力只产生了不足 8% 的 GDP。未来中国要持续提高生产力，参考目前英德法等国家的劳动力比例，到 2035 年，中国要将超过 20% 的农业劳动力转移到第二、三产业；每一产业内部也将会有更多的重新配置。考虑到 AI 应用，一方面它会造成大量失业，另一方面，我们也要看到希望，希望 AI 技术能够加速产业重新配置的进程。

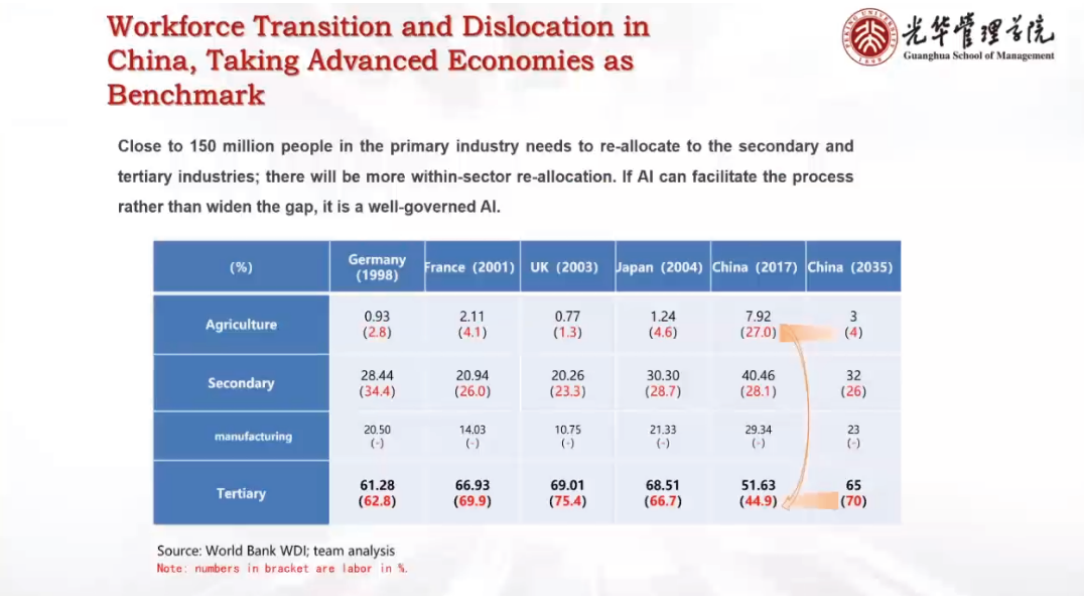


图 4：中国劳动力市场转移和错位

最后，我总结一下，我们要做到如下几点：

1. 充分意识到 AI 应用的意料之外的后果，愿意讨论伦理问题和治理问题，积极行动，在利益相关者之间达成共识；
2. 建立有效机制推动社区内外的讨论，使 AI 从业者意识到 AI 伦理和治理有关问题，寻找最佳实践并培养自律意识；
3. 学习美国经济学家鲍莫尔（William Jack Baumol）于 1967 年提出的“鲍莫尔成本病”（Baumol’s disease）的见解：经济增长未必受制于我们擅长的领域（例如自动化和 AI 应用），而是受制于基础的而又难以改善的部分；
4. AI 治理就是我们如何理解“基础的而又难以改善的部分”。一个尝试性建议是我们需要建立一个 AI 应用的约束模型。

徐云程：我将从公司层面来阐述，因为 AI 伦理产生约束力以及围绕这一主题的讨论真正落实到行动也离不开这一视角。接下来我会介绍旷视在 AI 伦理治理的行动与收获以及对于未来的考量。

图 5：确保 AI 的可持续发展

我先解释下可持续发展（Sustainability）的概念对旷视而言的重要性。拿旷视公司来说，可持续发展意味着两方面的内容，一是如何建立一个可持续发展的 AI 公司，进而发展可持续的 AI 产业；二是作为 AI 公司，我们在思考 AI 技术能如何为可持续发展社会略尽绵薄之力。

那么，我从旷视自己的故事讲起。大约是在一年前，我们意识到引入 AI 伦理并将其纳入公司发展策略是多么重要。这一契机源于公司创始人兼 CEO 印奇，他毕业于清华大学，依靠深度学习的专业技术背景，和其他两位清华大学毕业生于 2011 年共同创建了该公司。创始之初，AI 技术还没有达到现在的火爆程度。可以看出，他们是坚定的技术拥趸者，跳进了深度学习的技术领域，以计算机视觉为切入点开始探索。他们是技术造福社会的硬核支持者。一年前，他开始给我们传达 AI 伦理的必要性，认为公司要建立 AI 伦理原则。

一开始，这对于我来说是很新鲜的。我通过学习了解到，当时国际上已经有了很多关于 AI 伦理的讨论。得益于智源研究院等学术机构、曾毅这样的支持者以及其他国家的科技公司，我们很快建立了公司的 AI 伦理原则。第一步，在 2019 年 7 月，我们推出了自己的《人工智能应用准则》，涵盖了六个维度。这一准则其实是公司信仰的基石，不仅如此，身为务实的 AI 科技公司之一，我们要向外展示如何使准则落地实施。第二步，由于话题太新，我们很难写出执行计划。因此，我们需要一个机制，需要一个公司外部的智者来帮助我们。于是我们向董事会报告，成立了 AI 伦理委员会。除了公司内部员工，该委员会还有公司外部委员，他们能够提出不同的观点及行动指南，有时会挑明事情的本质并给予真诚反馈，工作复盘，调整走向，指出对错。我们意识到公司出台 AI 伦理治理原则需要听见不同的声音，这就是我们成立 AI 伦理委员会的初衷。第三步，我很自豪的是，在公司内部成立 AI 伦理管理委员会。这么做的理由是，尽管 AI 伦理委员会是高规格的，但公司要在数据库方面做到领先地位，在公司内部就要有管理团队来考虑准则的遵守与实施。我们一手提拔了部分符合要求的公司高管，组成了这一管理委员会，委员阵容包括从公司科研主管到 HR 主管。内容包括如何保护数据，如何训练模型以及告知公司员工遵守伦理原则做出编码产品，这些是非常有效的模式。公司外部委员及公司内部董事会管理委员形成管理架构，并进行有效运转。

今年年初，公司积极倡导 AI 伦理，试行 AI 治理模式，因为我们很快意识到工业界存在很多共通的问题亟待解决。作为科技公司，这些解决办法使用了数据库案例，公司有很多科学家来解决针对数据库的实际应用问题，

除此以外，也要与不同对象合作寻求答案，需要进行深度研究项目。这也是为何在年初我们宣称要建立 AI 治理研究院。该研究院也不单单服务于旷视公司，它表明了公司想要与研究项目的各方建立联系的态度。

以上是关于旷视公司的故事介绍，我们历时一年半来做这件事，刚才我解释了公司开展 AI 伦理治理的理由。接下来回答最开始提出的第一点，一个可持续发展的 AI 公司如何引领可持续发展产业。公司这样做不只是因为我们是理想主义者，更为重要的是，当我们回顾历史，技术进步的流程是先向前发展，然后遇到风险与问题的时候后退和治理。AI 业界为此付出了惨痛的代价，这对于业界和公司都是沉重的打击。因此，业界与公司要形成一种平行模式，就是说，在发展技术的同时，AI 伦理考量和 AI 治理也要携手并进。我们不允许 AI 伦理与治理处于滞后于技术的状态，这是旷视公司的态度，也是遵循的重要策略。尽管仍有很多问题需要面对，这一模式也要执行。“以身作则，言行一致”（Model the way, walk the talk）是我们公司奉行的原则。

以上是对于第一点的阐释，至于第二点，AI 公司能为社会可持续发展做点什么？我很感谢我们有很好的朋友，包括智源研究院和北大光华管理学院，有如此强大的领导力量，给予了旷视很好的建议来服务于构建可持续发展的社会。接下来我给出一些例子来说明。

年初，我和一位学者讨论，旷视公司在自身经历与实践以外，可以唤起人们的创新意识、推出意见领袖，在中国宣称 AI 治理是商业的新模式，可以实现很多可理解的工业发展，是可持续的快速发展道路。我们做了一些公司的 AI 伦理的个案分析，发布了《全球 AI 治理十大事件》的文章。我们在文中梳理清楚了一些典型事件，当然并不代表我们认同它们的处理方式是完全正确的。这是我们做的第一件事，第二件事是我和曾毅多次讨论过的，即如何回应公众的关切和疑问，但目前 AI 产业与公司之间没有一个现成的交互模式。我从新闻里读到了很多讯息，我知道哪些是真的，哪些是荒唐的。今天我听见不止一位演讲者提到透明度，透明对话才能建立信任。因此，我们要做的是告诉人们，没啥可隐藏的。年初，我们在中国与公众进行对话，其中一些是 AI 伦理的挑战者，我们将一些普遍的问题汇总成 10 个问题，并邀请观众分享自己的观点。令人吃惊的是，观点差别很大，记得有篇短文在 72 小时内有超两千万的点击量，有一千多人评论。这件事说明，一旦有了正确机制，且愿意与公众互动，他们就会和你交流想法。我们随后消化和整理了这些反馈，快速整合为四点公众关切的内容： 一是 AI 伦理与法律的本质在新兴领域的作用； 二是经济发展与 AI 技术发展背景下的社会公平； 三是问责机制和利益分配； 四是个人安全和隐私保护。这四点的共同目标可以驱动集聚经济、技术、哲学及决策者利益各方，这种团结一致也是我很欣赏智源研究院为可持续发展而努力的方面。可持续发展目标（SDGs）有 17 个方面，宗旨是为人类福祉而努力。现在大家已经达成了共识，AI 技术服务于可持续发展是正确的道路，将人们团结起来，这也是我们研究院加入可持续发展目标项目的原因。

总结一下，我们聚焦于两点，一是建立社区 AI 治理模式，这与联合国 SDGs 的第 11 条契合，因为大部分人生活在城市和社区。； 二是信任 AI，具体来说，是信任计算机视觉。作为负责任的 AI 公司，纳入 AI 伦理治理，将其内化为公司文化，形成和实施自己的治理模式，与不同的团体合作，建设可持续发展的未来。

最后，我讲一下机制，本场圆桌论坛主题是连接东西方 AI 伦理治理为 AI 的下个十年努力。虽然是中国人，但我是在美国接受的教育，也在美国公司工作多年，我认为我自身体现了中西方之间天然的连结。单看我们自己，我们生活在某个具体的国家，但从 AI 技术角度来看，东西方并无本质差异，我们聚集起来为了一个共同的目标，利用影响深远的 AI 技术为未来十年，为子孙后代，积极合作，不管是作为公司一员，还是身为市民。结合旷视和我自身经验，有三点很重要： 一是全球对话与交流，人们需要交流、分享和透明的信息。二是付诸实践，

许多事情要植根于应用，因为公司是接近市场和用户的。三是有很多团体对 AI 治理感兴趣，现在就是关注实际和严肃话题的时机，指出问题很容易，但解决问题很难，这需要东西方各个团体通力合作，人们不该避开这次挑战。

吴国斌：感谢曾毅的邀请。我会介绍滴滴相关内容以及我对 NGO（非政府组织）的热情。滴滴是国际领先的便捷移动出行平台，我们公司提供了基于移动应用的全方位交通和生活方式服务。滴滴在亚洲、拉丁美洲、澳洲等地共有 5.5 亿用户，平台每年提供出行 100 亿旅次，可以说是排名第一。

说到此次论坛主题，我想讲下滴滴为实现 SDGs 所做的努力和贡献，它的 AI 项目实施。滴滴是为缓解交通和环境压力而存在的，我们借助 AI 技术能力把智能出行进行本地化创新。例如，滴滴利用 AI 和大数据技术设计智能出行方案，包括交通出行大数据平台、智慧信号灯、安全驾驶分析等。我们缓解了交通拥堵，提升了道路安全，促进了可持续的居民社区建设。滴滴采用大数据技术和 AI 算法来设计可预测调度，规划路径，计算最优上车点，减少空驶现象，发展共享出行，叫拼车业务（Car pooling business）。该业务可以提高资源利用率，减少二氧化碳（CO2）排放量，从 2018 年到 2019 年，滴滴 CO2 减排量达到 130.3 万吨，相当于 68 万车辆一年的排放量（编者注：口误说成 1300 万吨，已据《2020 滴滴平台绿色出行白皮书》校正）。

2018 年，滴滴开始探索利用 AI 技术平台为社会谋福祉，与数十家高校、研究机构和社会组织合作，核心研究方向包括安全性（Safety，司机安全、司机健康和环境状况），移动出行（Mobility）和辅助功能（Accessibility），例如选择优良空气质量地图以解决环境问题，优化新能源整合方案和改进智能出行技术，以及司机智能助手介绍。另一例子是滴滴推出的盖亚数据开放计划（GAIA Open Dataset，网址：https://outreach.didichuxing.com/research/opendata/），依托领先的大数据和技术优势，面向学术界提供真实的脱敏数据资源，开放协作，旨在以产学研深度融合推进交通领域的基础性与前瞻性研究和成果转化，提速智慧交通领域的科研发展，为社会发展创造更大价值。迄今为止，我们有开放的轨迹数据集（Trajectory dataset）、POI 检索数据（POI retrieval dataset）和大规模行车视频数据集（The large-scale driving video dataset）。

最后，实现 SDGs 需要创造性与创新性的参与，同时 SDGs 也绘制了发展蓝图，为企业发展模式提供机遇、创新和社会责任实践。此次圆桌论坛举办之前，“新一代人工智能和可持续发展目标（AI for SDGs）”研究项目已经启动，滴滴非常愿意支持此项目，与其他机构一起推动 AI 创新，实现全球可持续发展，滴滴也有意支持交通出行改革、可持续发展城市、人类福祉等研究项目，以及公众关切问题，比如利用 AI 技术做一些更有意义和价值的事情，基于 AI 伦理原则实现人类和谐、远离伤害、实现公平和肩负责任感。谢谢。

曾毅：非常感谢。接下来我们进行互动讨论，Vicent（编者注：Vicent Muller，埃因霍芬理工大学教授，IEEE 机器人与自动化学会机器人伦理委员会共同主席）提出了一个有意思的问题，我觉得是向刘俏提问的，请 Vicent 将问题复述一下可以吗？

Vicent Muller：首先感谢本次会议，受益良多，我想提的一个的问题是：中国是否从负面消极的教训中汲取了经验呢？中国发展迅速，不断追赶美国或其他国家的技术水平，但我作为西方一派，并不觉得我们所取得的成就有多耀眼，发展的曲线进程中也会犯错，我认为中国也会遇见同样的问题。我儿时生活在德国的小地方，深受钢铁厂等重工业污染影响，中国十几年后或许也会面临类似问题。所以要总结教训，改正错误，减少环境污染，避免社会问题，中国避开我们的失败可以在以后少走弯路。

刘俏： 这是个好问题。中国从美国和其他发达经济体学到的是做好财富分配。人们很大程度上认为，全球化经济发展目标不能惠及所有人，中国确实有很多人还未享受到经济发展带来的福利。以美国为例，法国经济学家托马斯·皮凯蒂（Thomas Piketty）系统评估了美国 1978–2015 年间的财富分配，发现占社会人口 40% 的中产阶级人均 GDP 在过去四十年里增长不足 1%，证明这部分人并未共享经济发展的成果，资本过度追求效率且不关心社会问题。人们讨论追逐美国梦的过程中出现的财富公平问题，所以中国要在发展过程中警惕分配不公问题。我们讨论了 AI 技术应用，同时也要思考如何避免重蹈覆辙，做好财富分配，而这些也属于可持续发展和 AI 伦理治理的内容。

Vicent Muller： 财富分配确实是一个事关 AI 技术的重要问题。工业发展不依赖大量资本涌入或占据某个具体地段，而是人们使用产品，然后供应方一夜暴富，AI 技术会加速这一进程。您所说的情况在欧美有所不同。

曾毅： 谢谢。从 Vicent 的提问中，我们可以借鉴西方的发展历史。我想分享一个我喜欢的近期发生的故事，讲讲东方可供参考的做法。大约一年前，有些公司想在中国大学教室应用面部情绪识别系统，结果大部分师生都不同意，教育部也叫停该了行动，以防发生不好的事情。不出所料，西方媒体对此批评不断。而在之后不久，卡耐基梅隆大学（CMU）推出了一款自认为酷炫的应用（注：指的是 CMU 推出的 OpenPose 人体姿态识别项目），除了可以识别面部表情，还可以识别课堂上的手部运动。对于技术革新者而言，他们认为这是很棒的尝试。同样的故事才刚刚在中国上演过，却有评论写道或许中国并不在意 AI 伦理，这显然是有失偏颇的。其实在中国大家也不太支持这样做，又怎么能指望西方社会接受呢？我想表达的是，西方学习的这个东方故事，或许在东方自己的文化里并不被接纳。出于各种考量，这样的尝试或许应该暂时搁置。或许这就是东西方相互学习的原因吧。

现在有两千多位观众参与这场视频会议，我们来挑选一个观众问题进行回答。我将这个中文问题翻译一下：现在 AI 分析师可以识别语音、人脸、指纹以及手写字体，技术获取这些数据相当容易，尤其在疫情期间，我们必须居家工作，用 Zoom 或者腾讯会议线上办公，想要获取我们的数据，尤其是人脸数据，就更简单了，或许此时我们的面部数据就正在被收集。现在人类如何战胜疫情，如何应对 AI 获取更多的个人隐私数据？是否有更好的方式约束可获取的信息？尤其在疫情期间，我们怎样保护自己的个人信息不被泄露？现在视频会议很流行，初衷是提供交流互动平台，但现在也引人担忧。有没有可以在启动视频会议时保护隐私数据的技术呢？ 请大家回答一下这个问题吧，有请 Vicent。

Vicent Muller： 我认为这个问题很重要。尤其是在获得政策力量支持的情况下，现在是否存在用户满意的数据使用技术呢？例如，中国政府及其他政府的软件加工，是对国家的贡献。我今年开设的数据科学伦理课程中讨论了报告新冠疫情的移动应用，这些应用需要的数据或多或少。遵循只获取运行所需数据的原则，就能做成一个好的系统，比如连接蓝牙时不获取设备位置信息和存储信息。英国人民也很反感存储信息和通话记录等被收集，民众的反对会导致政策损失甚至经济损失，而这源于公司技术产品存在不合理收集用户数据的行为。我们现在使用的 Zoom 视频会议软件在疫情期间风靡欧洲，但是我所在的大学明确告知师生不支持使用 Zoom，因为校方认为该软件存在隐私风险。确保数据安全需要共同努力。

Seán Ó hÉigeartaigh： 我同意 Vicent 所说的，这是很重要的 AI 技术涉及过度收集数据的伦理问题。目前在英国与此相关的讨论很多，而 Vicent 提出的观点很重要，究竟只获取应用运行所需数据的界限在哪里？数据存储期限是多久？数据是如何存储的？怎样确保数据安全？是谁在怎样的情况下有接触数据的权限？举例来说，我

很愿意将我的信息提供给 NHS（National Health Service，英国国家医疗服务体系），我想知道我的个人数据会被用来做什么。我想强调的是，就算部分信息匿名处理，但将各个软件获取的数据拼凑起来也能得到超出我们本来预期的个人信息，我们要做好权衡。比如在英国追踪新冠患者密切接触者的过程中，以上信息就很有用，可以集中关注患者活动轨迹范围，了解英国疫情传播情况，这有助于拯救生命，考虑到伦理问题，疫情传播会加重医疗服务人员的负担，所以要综合实际情况进行考虑，在危急时刻做出正确决策，收集更多个人信息，而在疫情缓解之后停止这样的行为。正如 Vicent 所言，在线会议平台在用户使用时要求提供个人数据，还收集信息和视频数据。用户较多的软件系统如果有漏洞，可能会发生恶意攻击性犯罪活动，我们要警惕系统的疏忽。近年来在英国对计算机系统批评的声音很多，我们要对信息安全提高警惕。

徐云程：以上二位发言给人很多启发，我想从业界和公司的角度讲三点个人观察： 首先，应对更大的话题和挑战需要群策群力。一是决策者要考虑监管问题，这是重大决策。监管不明的情况下，公司该要如何指导行动呢？比如鼓励公司自我治理，不止一家公司提出了自己的数据隐私保护原则，这在东西方都越来越普遍。二是经济机制。刘俏院长在这方面就做得很好，当意识到个人数据也有价值的时候，就像个人财产不容侵犯一样，人们会更加保护个人信息，这是一种经济鼓励的动力。三是技术本身会寻找技术挑战的解决办法。业界公司会尽力在节约成本的情况下使用最先进的技术，我认为需要结合各种办法来解决问题。其次，面部识别技术很火，有很多相关讨论，可以利用这项技术找回走失儿童。所以技术应用要设定使用场景，有了具体场景，讨论才有意义，才是公平地看待技术应用，才能去思考数据存储周期等问题，要看重使用的目的。我们要对此更加具体和严谨。最后，“不知为不知”。AI 技术处于商业化的早期阶段，我们要有敏捷的思维方式，保持学习，不断思考，及时调整。边做边学，先做后学，要有互联网思维，来有效应对这个宏大话题的未知与未来。

曾毅：谢谢。今天我们从东西方视角讨论了关于 AI 治理与可持续发展的现在与未来。实现目标的方法多种多样，在场的各位也付出了很多努力。科学家和业界从业者有共同的目标，不仅考虑政策因素，还有我们对未来的责任感，全球政策合作非常重要，经济组织和工业界要共同努力创造全人类的美好未来。感谢各位的参与和努力，我相信这只是一个起点。欢迎各位与智源研究院共同见证面向可持续发展的人工智能智库（AI4SDGs Think Tank）的成立，以及面向可持续发展的人工智能公益研究计划（AI4SDGs Research Program）的发布。