CCID 赛迪智库 译丛

2022 年 6 月 13 日　　第 23 期　　总第 542 期

建立标准，发现和管控人工智能存在的偏差

【译者按】2022 年 3 月，美国国家标准与技术研究所（NIST）发布《建立标准，发现和管控人工智能存在的偏差》报告。报告认为，人工智能中的偏差问题会对个人、组织和社会产生一系列负面影响，需要采取社会技术的系统方法加以应对。报告介绍了人工智能偏差的概念，分析了由此产生的各类危害与挑战，并建议从数据集、测试评估验证环节、人为因素三个关键维度制定初步的人工智能治理社会技术框架，进而提出了相应的操作指南。赛迪智库信息化与软件产业研究所对报告进行了编译，期望对我国有关部门有所帮助。

【关键词】人工智能偏差 标准 治理

随着人工智能（AI）系统更多参与跨行业及关键领域应用，其技术流程中普遍存在的偏差问题可能会造成有害影响，这给社会公平及 AI 系统的公众信任埋下了隐患。然而，当前社会对于人工智能偏差的认知尚不充分，应对人工智能偏差有害影响的尝试仍然集中在计算性因素上，比如数据集的代表性和机器学习算法的公平性。这类补救措施对于减少偏差至关重要，但还远远不够。人为因素、系统性的制度因素以及社会因素也是人工智能偏差的重要来源，但目前却未被重视。要成功应对人工智能偏差的挑战，就需要考虑所有形式的偏差。为此，本文介绍了人工智能偏差的概念及分类，讨论了偏差产生的原因及带来的挑战，并从数据集、测试评估验证环节和人为因素三个方面为制定详尽的社会技术指导路线提供了初步指南。

一、人工智能偏差:背景和术语

（一）人工智能偏差相关概念

1、人工智能偏差的定义

统计性定义：在技术系统中，偏差通常都被理解为一种统计现象。与随机误差不同，偏差是一种通过对统计结果进行系统性扭曲从而破坏其代表性的效应。国际标准化组织（ISO）将偏差更广泛地定义为：“参考值偏离事实的程度”。因此，当 AI 系统表现出系统性的失准行为时，就可被认定存在偏差。这种统计

性视角并未充分涵盖或揭示 AI 系统中存在偏差所造成的全部风险。

法律性定义：对人工智能偏差的讨论不能脱离美国法律体系中针对偏差的处理办法，以及偏差与解决歧视和公平性的法律法规之间的关系。目前，对于不允许的歧视性偏差，法院一般会采取差别对待或差异性影响两种方式进行定义。监管机构与法院尚没有统一的办法来衡量所有不允许的偏差。

认知和社会背景：人工智能系统设计和开发的团队将他们的认知偏差带入流程，致使偏差普遍存在于各项假设中。若系统性偏差存在于制度层面，则会影响到机构或团队的结构和决策流程的掌控者，带来人工智能生命周期中的个人和群体启发性偏差与认知/感知偏差。同时，终端用户、下游决策者和政策制定者做出的决策也会受到这些偏差的影响。由于影响人类决策的偏差通常是隐性且无意识的，因此无法轻易地通过人为控制或意识纠正进行限制。

2、人工智能偏差的类别

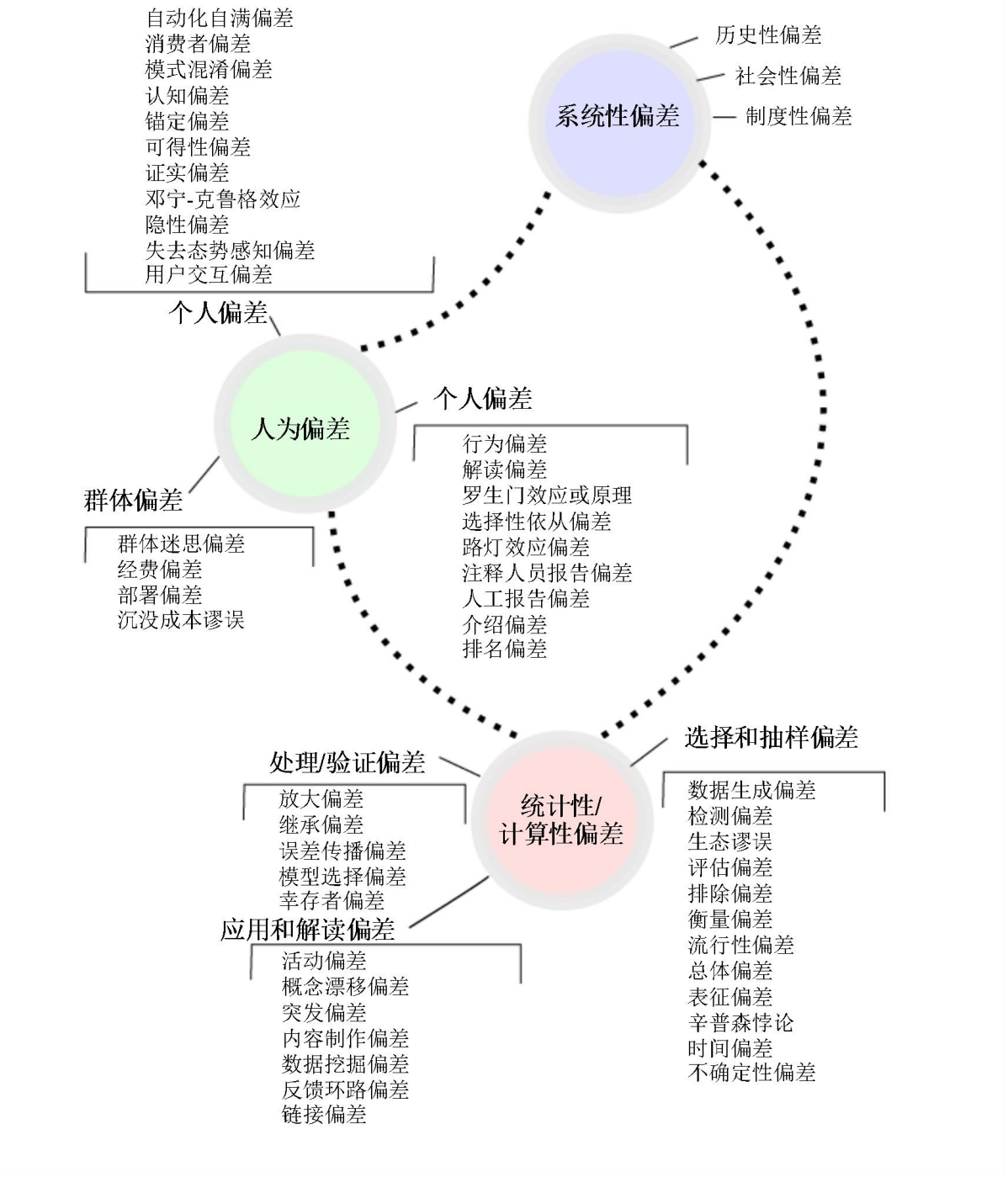

图 1：人工智能偏差的类别

系统性偏差：系统性偏差也被称为制度性偏差或历史性偏差，源自特定机构的程序或做法，其运作方式致使某些社会群体处于优势地位或受到青睐，而其他社会群体则处于劣势地位或受到贬抑，如制度性种族主义和性别歧视。这些偏差来源于人工智能使用的数据集，乃至贯穿人工智能生命周期，存在于更广泛的社会制度规范和流程中。

统计性和计算性偏差：统计性和计算性偏差源自样本不能代表总体所导致的误差。这些偏差由系统性错误而非随机性错误所导致，而且在没有偏见、偏袒或歧视意图的情况下也可能发生。这些偏差存在于开发人工智能应用所使用的数据集和算法过程中，当算法针对某一类型的数据进行训练且无法进行外延时，偏差就会产生。

人为偏差：人为偏差反映的是人类思维中的系统性误差，这些误差源于启发性原理数量有限以及基于简单判断进行数据预测。人为偏差往往是隐性的，而且很可能与个人或群体如何感知信息以进行决策或填补缺失或未知信息有关，仅仅提高对偏差的认识并不能确保对它的限制。这类偏差无处不在，贯穿人工智能生命周期中的机构、群体和个人决策过程，乃至人工智能应用部署后的使用过程。

（二）人工智能偏差的危害

一方面，当利用人工智能提供决策支持时，若没有人工操作员对其进行相关约束，机器学习模型常常会由于“认知不确定性”和“偶然不确定性”等影响而造成糟糕表现。而且目前用来捕捉这些模型的有害影响及其他后果的方法既不精准也不全面。

另一方面，机器学习系统能否依照人类社会的价值观进行学习和操作仍是一个亟待研究和关注的领域。系统性偏差和隐性偏差可能通过训练时使用的数据，以及支撑人工智能委托、开发、部署和使用方式的制度安排与做法而带入。同时，统计/算法偏差以及人为偏差存在于工程与建模过程本身，而无法正确验证模型性能使这些偏差在部署过程中暴露无遗。这些偏差与个人的认知偏差相冲突，若不加以应对，可能会形成一个复杂而有害的混合体，对个人和社会造成远超传统歧视性做法的负面影响。

（三）应对人工智能偏差危害的新视角：社会技术视角

传统堆叠技术解决方案并不能充分反映人工智能系统的社会影响，仅从计算角度试图解决偏差存在局限性。因此，要将人工智能扩展到公共生活的方方面面，需要将人们的视角从纯技术角度拓展为实质上的社会技术视角，站在更宏大的社会制度层面来思考人工智能。

通过社会技术视角来重新构建与人工智能相关的各项因素，

具体包括三个维度：数据集；测试评估、确认及验证（TEVV）[1]；参与式设计及“人在回路”等人为因素。以上几个维度可以更全面地理解人工智能的影响和贯穿其生命周期的种种关键决策，并实现偏差的动态评估、了解影响偏差大小的条件及偏差间相互作用的机制。同时，实现个人、群体和社会需求的兼顾，还需要广泛的学科和各相关方充分参与。

（四）更新后的人工智能生命周期

为了使人工智能相关技术人员将人工智能生命周期过程与人工智能偏差类别联系起来，有效促进对偏差的发现和管控，本文给出了一个四阶段人工智能生命周期（图 2）。

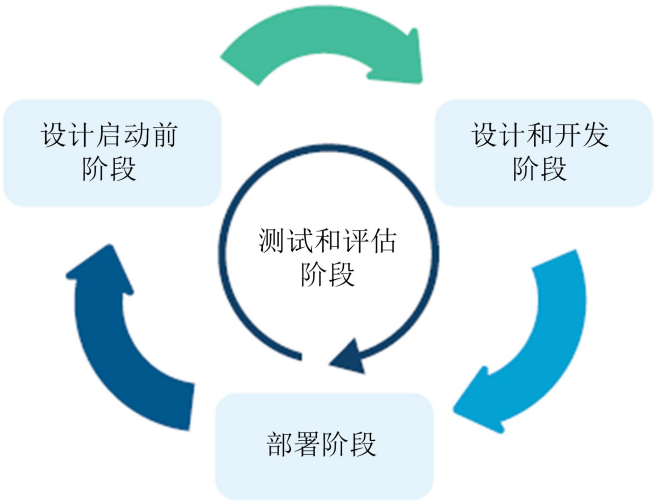

图 2：人工智能开发生命周期

设计启动前阶段：该阶段主要是进行规划、问题说明、背景研究和数据识别。此阶段核心在于确定有话语权或控制权的个人

[1] 人工智能开发生命周期中的部分环节。

或团队来进行相关问题的决策。这些早期决策及其决策者可以反映出机构环境中的系统性偏差。此外，系统性偏差也反映在设计启动前所选择的数据集上。所有这些偏差都会以复杂的形式影响后期阶段和决策，并导致结果的偏差。

设计和开发阶段：该阶段通常从分析要求和可用数据开始，并以此为基础进行模型设计或选择。在设计过程中，应当通过兼容性分析找出潜在的偏差来源，并评估和调整偏差应对措施。在开发过程中，机构应定期评估偏差，发现流程的完整性及应对措施的有效性。最后，在开发阶段结束、正式部署之前，有必要对偏差应对措施进行全面评估，以确保系统保持在预先设定的范围。在模型正式发布和部署前，其总体模型规格必须包括已被确定的偏差来源、已实施的应对技术以及相关的性能评估。

部署阶段：该阶段是 AI 系统发布和使用阶段。技术团队应当实施持续监控，并制定详细的策略和程序来处置系统的结果和行为。可能需要对系统进行重新训练以纠正其副作用，甚至关停系统，以确保其应用不会造成非预期的影响或危害。

测试和评估阶段：该阶段贯穿整个人工智能开发生命周期。此阶段鼓励所有机构对一切可能受到偏差影响的 AI 系统组件及功能进行持续测试和评估，以确保评估的平衡性与全面性。如果得到的结果不符合预期，则应将其反馈到模型的设计启动前阶

段，对模型设计的任何拟议变更均应与新的数据和要求一起接受评估，然后开始新一轮的设计和开发，确保此前发现的所有问题均得到解决。

二、减轻人工智能偏差面临的挑战与建议

（一）人工智能偏差中的数据集

1、数据集方面存在的挑战

人工智能的设计和开发高度依赖大规模数据集，这种需求可能会引导研究人员、开发人员和从业人员更在乎数据集的可用性或可得性，而无论其合适与否。结果是，当现成的、却不能完全代表目标总体样本的数据集被反复用作训练数据时，系统性偏差也可能会表现为可得性偏差[2]。同时，即使数据集有代表性，也仍可能表现出历史性偏差和系统性偏差。由于受保护属性的隐藏信息可以通过代理或潜在变量推导出来，揭露出个人和群体的非必要信息，因此基于这些变量的模型仍然会对个人或某一类人造成负面影响，可能会造成歧视。

当终端用户与 AI 系统发生交互时，这些不当或不完善的早期设计与开发决策使得该过程容易受到额外的统计性偏差或人为偏差的影响。例如，算法模型可能仅建立在最活跃用户的数据之上，其创建的后续系统活动可能也并不反映目标或真实用户群

[2] 指人们往往根据认知上的易得性来判断事件的可能性，而忽视对其他信息的关注进行深度发掘，从而造成判断的偏差。

体。此外，反馈环路可能会将误差进一步放大，使得随后的训练数据全部来自于最活跃用户，进而将造成潜在的有害影响。

2、数据集方面的改进建议

应对统计性偏差：应对人工智能偏差的一个主要趋势是关注建模过程中所使用数据集的全面统计特征。对于算法模型来说，常见的算法技术都假设变量是单峰的。然而，数据却往往是异构和多峰的。因此，无论模型是用于基准测试、预测还是分类，必须记录和交流人工智能结果的适用性存在的局限。此外，在数据集的迁移使用时还需特别注意数据集分布中的潜在差异，并关注其对模型的不公平性与误差产生的影响。

应用社会技术方法：人工智能建模需要结合地区具体地理特征，因此，需要对机器学习应用中数据集的使用加以调整，以适应其部署环境中的所有社会技术因素。在设计阶段，社会技术分析带来了对某一现象的动态或特征性社会异变的深刻理解。这有助于更好地制定出问题分析框架，并对数据集合适与否做出评估。开发阶段的社会技术视角有助于选择数据源和属性，并明确将影响评估作为算法准确性的补充。

关注人为因素与数据集的相互作用：构建人工智能应用基础模型时，设计和开发团队关于使用哪些数据集的决定和假设会加剧数据集中存在的系统性、制度性偏差。同时，在数据选择、管

理、准备和分析过程中，人为偏差也会造成一定影响。例如，注释训练数据的人员可能会带入其自身的认知偏差；相关人员清洗数据源与变量时也会按照自己的理念行事；数据分析决策在边缘化总体样本中存在收集偏差。以上人工智能偏差和公平性的问题需要解决。此外，需记录人为偏差的潜在来源，以提升人工智能模型描述的透明度和可解读性。

（二）对人工智能偏差进行测试评估、确认及验证时的注意事项

1、TEVV（测试评估、确认及验证）方面存在的挑战

机器学习过程中的预测不确定性：机器学习存在两种类型的预测不确定性：“认知不确定性”和“偶然不确定性”。“认知不确定性”常在参数计算中出现。由于数学问题上的解值具有非唯一性，当真实数据与训练数据的分布不匹配时，可能会影响已部署的深度学习系统的行为，导致有害偏差。“偶然不确定性”代表数据中固有的不确定性，是预测不确定性中不可再分的部分。例如，训练数据集的标签分配过程中的不确定性。

大型语言模型的发展带来挑战：大型语言模型在深度学习中的重要性不断增加，但其在“认知不确定性”和“偶然不确定性”方面造成了重大挑战。依赖大量未经整理的网络数据会增加偶然不确定性。

模型设计与数据处理流程的偏差问题：为了让 AI 系统确认建模的侧重点，技术人员在对数据进行分类和排序时往往会将背景信息扁平化处理，并对不可观察现象进行量化处理，这一操作可能会导致有害偏差。同时，软件设计师和数据科学家对系统性能进行优化的过程，也可能会在无意中成为人工智能系统偏差的来源。此外，在模型选择过程中忽略背景信息也可能会导致子群体的结果有偏差。相应地，使用群体汇总数据预测个人行为的系统可能会导致结果出现偏差。这些无意中对某些因素加权后得出的算法结果，可能会加剧和固化社会的不平等。

算法复杂性的偏差问题：出于成本与实现难度的考虑，技术人员所使用的通常都是参数较少的简单模型。然而，这类模型对训练数据的限制性假设通常不兼容有细微差别的统计资料，可能会加剧统计性偏差。复杂模型通常用于文本图像等非线性、多模态数据，这种模型捕捉的潜在的系统性偏差可能难以识别和预测。

系统验证的有效性问题：在系统验证环节可能会出现许多困难和缺陷。系统测试往往缺少真实数据、或者噪声标签及其他注释因素，这导致人们很难知道什么是准确的。同时，代理变量的使用再度加剧了这一困难。在最佳条件进行系统测试，是另一项极具挑战性的设计缺陷。同时，系统性能指标难以概括，也可能

导致非预期使用的问题。

验证和部署环节的偏差问题：验证意味着确保系统不会以非预期的方式使用。当人工智能模型以开发人员不希望的方式使用时，就会产生部署偏差。当部署的系统存在缺陷或疏漏时，这类系统会损害用户利益，并可能违背现行的法律框架，加剧公众对人工智能技术的不信任。

人工智能系统的“黑箱”问题：人工智能在黑箱可解释性、再现性问题和试错流程方面表面上看似科学，实际上难以遵循假设可检验、实验可解释、模型可证伪的科学方法。除此之外，高水准的机器学习库和低成本的云计算使人工智能的开发正变得越来越普遍。然而，人工智能本身在很大程度上仍不透明，深度神经网络和贝叶斯推理需要高等数学才能理解。

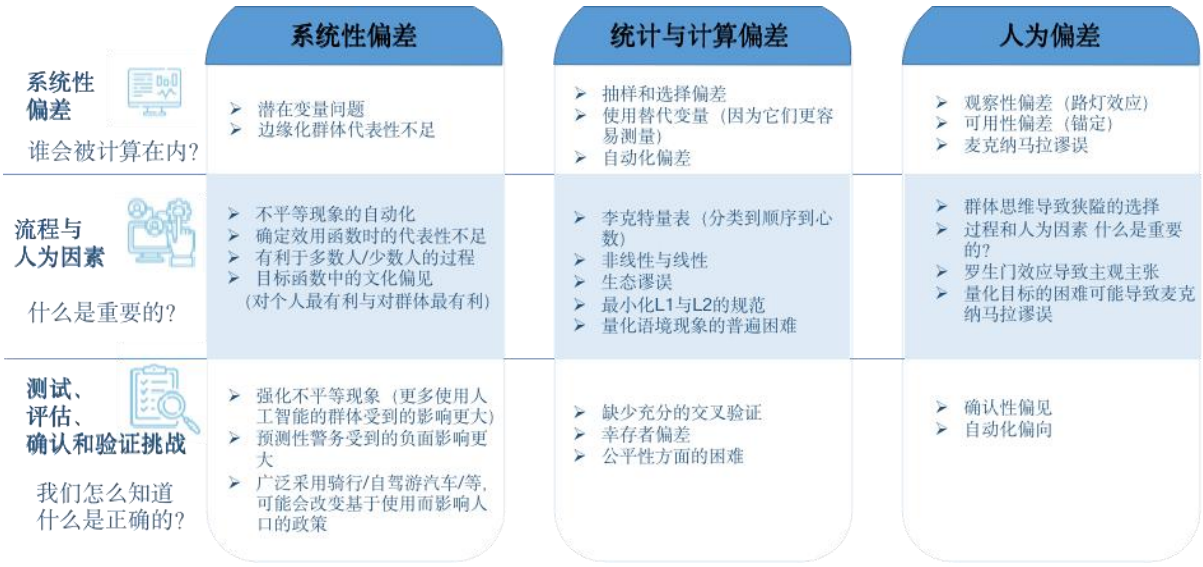

图 3：偏差是如何造成危害的

2、TEVV（测试评估、确认及验证）方面的改进建议

减少算法偏差：在机器学习中，若缺少算法应用于具体任务的背景信息，给模型或算法指定偏差本身是没有意义的。例如，在自然语言处理的背景中，仇恨言论检测模型使用方言标记作为毒性预测因子，这可能导致对少数民族群体的偏差。根据在特定任务背景下减少算法偏差的不同方法，目前的除偏方法可归为以下三类：预处理、处理、后处理三类。它们分别通过数据转换、算法修改、黑箱模型推导的方式，减少训练过程中的偏差问题。

完善公平性指标：目前的研究表明，公平性简化成一个简明的数学定义是困难的，同时观察性的公平性指标尚有待发展。公平性是动态的、社会性的、特定于应用和环境的，而不仅是一个抽象或普遍的统计问题。因此，必须采用社会技术方法实现公平，以便为不同的环境提供现实的公平性定义，并为机器学习模型的开发和评估提供特定任务的数据集。

（三）人工智能偏差中的人为因素

1、人为因素方面存在的挑战

设计与开发环节的实验环境与现实存在差距：AI 系统的设计和开发是为了在特定的现实世界环境中使用，但往往在理想化的场景中进行测试。一经部署，最初的意图、理念或影响评估都可能发生偏移。不同的部署环境意味着需要考虑一系列新的风险。

在决定构建 AI 系统之前，需先与可能受到这些技术部署影响的广大相关方群体进行接触，这一点至关重要。

“人在回路”系统产生相关干扰：大多数算法决策系统都属于社会技术系统。从机器学习过程所使用的数据集和构建人员所做的决策，到与提供见解和监督保障系统运转人员的交互，这些过程都与人类社会行为密不可分。通常认为将人类置于 AI 系统的“回路”中能够杜绝不良事件的发生，但当前对于人工智能中“人在回路”的作用和责任的看法尚不明确，对这类系统性能的预期也往往基于未经检验的假设。对于专家推动型机器学习，专家可能会与机器学习模型进行交互，但较少参与系统本身的设计或开发。例如，在医学领域的 AI 系统中，医学方面的专家不一定熟悉机器学习、数据科学、计算机科学或其他传统上与人工智能设计或开发相关的领域。同时，行业专家倾向于借助 AI 系统来分担自身主观判断的压力，从而实现假定的自动化客观性，可能在无意中做出不准确和有害的决定。相应地，人工智能开发者群体可能会下意识地认为专家的方法已经得到了比实际情况更好的验证。这些隐性的个人和群体行为可能会创造条件，间接鼓励特别是高风险环境下对不完善技术的使用。因此，在没有评估和管理以上风险的情况下，专家推动型机器学习和“人在回路”的做法无法作为对 AI 系统及其结果实现有效监督的手段。

2、人为因素方面的改进建议

引入影响评估过程：一种名为算法影响评估的方法旨在确保以合乎伦理和负责任的方式进行人工智能技术开发。其中，发现和应对潜在的偏差是评估过程中的重要步骤。算法影响评估提供了一个高级结构，使机构能够框定每种算法或每次部署的风险，同时考虑到每个用例的具体情况。同时，参与影响评估还可以作为一种强制机制，迫使有关机构主动阐明一切风险，然后在发生任何危害时生成相应的缓解措施文件。此外，影响评估应是一个长期、迭代的任务，必须以合理的节奏反复进行影响评估。

推动相关方广泛参与：让终端用户、从业人员、行业专家，以及法律和社会科学界的跨学科专业人士等各种利益相关者参与进来，使得相关方可以利用自身不同的经验，拓宽 AI 系统工程师和设计师的视野，更全面地对 AI 技术应用的社会影响加以评估。同时，这种参与也需要审慎的规划和指导，以便与时俱进，改进实践。

提高团队多样性、公平性与包容性：确保参与训练、测试和部署系统的人员具备多元化的经验、专业知识和背景是一项关键的风险缓解措施，可帮助机构管控人工智能潜在的危害。为了将多样性、公平性和包容性的好处扩大到 AI 系统的用户和开发人员，应吸纳反映不同背景、观点和专业知识的个人进来，并帮助

机构更好理解系统的用户影响机制、系统对不同用户的利弊影响等问题。

运用生命周期法推进实践改进：通过采用生命周期方法，可以确定一些重要节点，开发完善的指导、保证和治理流程，并协助业务部门、数据和社会科学家对流程进行整合，以减少偏差。例如，在开发算法决策系统时列举制度性假设，并将这些假设与受影响的群体的预期进行对应。“有效文化挑战”的做法旨在营造一种环境，使技术开发人员能够主动挑战和质疑建模和工程中的步骤，从而帮助根除统计性偏差和人类决策中固有的偏差。

探索人机交互配置：为抵消已知的人类主观性和人为偏差，而将 AI 系统布置在高风险环境中这一做法仍然存在相当多的问题。必须有一种顾及到广泛的社会技术因素的“人在回路”方法，将人为因素、心理学、组织行为以及人机交互等领域与技术群体之间架起沟通桥梁。因此，应制定关于如何实施“人在回路”流程的正式指导，以避免放大或加剧可能在复杂环境下妨害结果的人为偏差、系统性偏差和计算性偏差。此外还需确定系统配置及其组件的必要标准，从而产生准确和值得信赖的结果。

提高系统和程序透明度：AI 系统需要更多的可解释性和可解读性。数据表和模型卡等工具的激增可以帮助填补这一缺口。同时，对于偏差与透明度相互交织的问题，透明度工具对于解决非

预期使用的问题带来帮助，但同时需要注意人类对人工智能模型结果的解读方式存在明显的个体差异，如专家与系统设计人员的问题关注角度不同。因此，协调指导是必要的，以确保透明度工具能为使用这些系统的专业人员提供有效支持，而不是间接放大偏差。

坚持以人为本的人工智能设计：以人为本的系统技术设计开发方法关注用户及其需求，全程以用户为核心，涉及硬件设计和软件设计等技术的所有方面，通过持续迭代不断改善整个系统的用户体验。以人为本的设计会更深入地关注公平性、偏差、价值观和伦理等重要的社会因素，本身也可作为其他开发生命周期的一部分，包括瀑布式、螺旋型和敏捷模式。

根据国际标准化组织（ISO）9241-210：2019 标准中的定义（图 4），以人为本的设计包括：对用户、任务和使用环境的明确理解，用户全程参与设计和开发过程，通过以人为本的评估来推动和完善设计，对原型进行设计、测试和修改的迭代过程，优化整体用户体验，具备多学科技术和观点的设计团队。

基于以上标准，用于开发 AI 系统的以人为本的设计方法越来越多地涵盖以下内容（图 5）：定义使用环境，包括操作环境、用户特征、任务和社会环境；确定用户和组织要求，包括业务要求、用户要求和技术要求；开发设计解决方案，包括系统设计、

用户界面和训练材料；进行评估，包括可用性和一致性测试。

其中，使用环境对 AI 系统是关键的社会技术考虑。在任何项目的前期，都必须考虑人工智能系统所使用的社会技术的动态和条件，以确保系统的设计能够满足用户需求、机构目标，以及更大的社会需求。使用环境不仅包括用户自身的使用环境，还涉及更广泛的背景状况，如开发人工智能系统的机构环境、即将使用该系统的操作环境、以及即将实施该系统的更大的社会环境。社会技术方法可以同时考虑技术层面和部署这些系统的复杂社会环境，提高关键的公众理解和能动性。

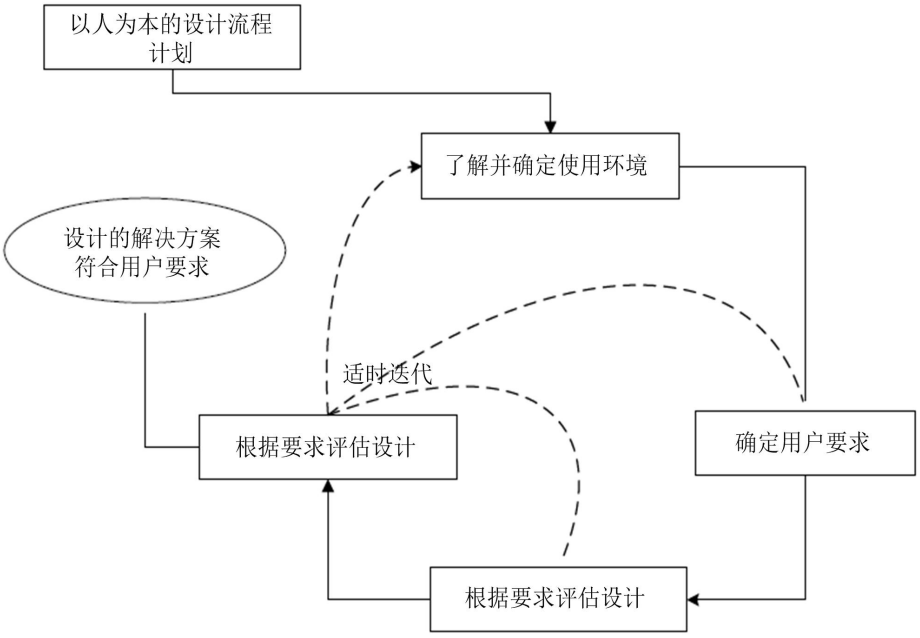

图 4：以人为本的设计流程【国际标准化组织 ISO 9241-210:2019】

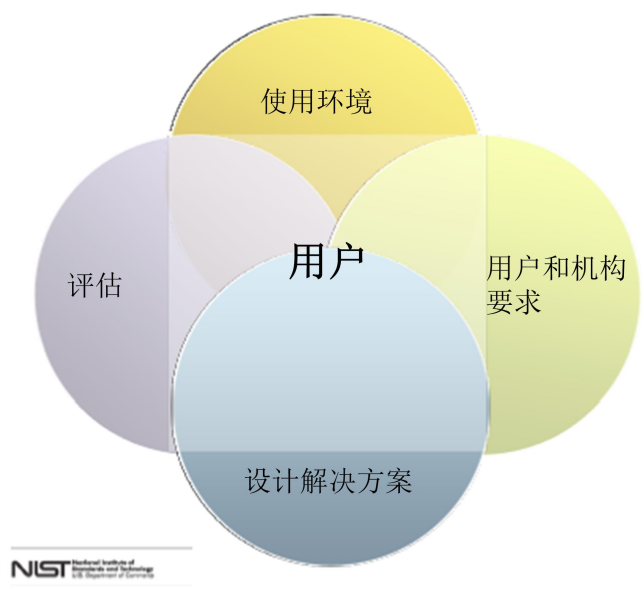

图 5：以人为本的 AI 系统设计流程

（四）人工智能偏差的治理指南

1、部署监测系统，预测潜在风险

部署额外的系统来监测潜在的偏差问题，可以帮助控制 AI 系统部署后表现与预测情况不同的风险，以便在检测到潜在问题时，及时提醒相关人员。

2、嵌入追诉渠道，优化事故补救

反馈渠道的可用性允许系统终端用户标记不正确或潜在有害的结果，并寻求对错误或危害的追诉。在人工智能系统中嵌入此类流程和技术，使得用户可以对错误的决定（甚至建议）提出申诉，同时也使技术开发团队能够在潜在事故的起始点对其进行补救。

3、完善政策与程序，加强风险管控

在 AI 系统的背景下，确保书面政策和程序能应对人工智能模型生命周期所有阶段的关键角色、职责和流程，对于管控和检测 AI 系统性能的潜在整体问题至关重要。政策和程序可以实现一致的开发和测试，有助于确保 AI 系统的结果是可复现的，使得相关风险可以得到一致的映射、衡量和管控。

专栏 1

如果在不同系统中使用了不可调和差异化指标，有关政策应：

- 定义与 AI 系统相关的关键术语、概念及其预期影响范围；
- 解决敏感或其他潜在风险数据的使用问题；
- 详细说明实验设计、数据质量和模型训练的标准；
- 概述应当如何映射和衡量偏差风险，及其依据标准；
- 详细说明模型测试和验证的过程；
- 详细说明法律或风险职能部门的审查流程；
- 规定持续审计和审查的周期和深度；
- 概述变更管控的要求；
- 详细说明与此类系统的事故响应相关的所有计划，以防在

部署期间出现任何重大风险。

4、建立标准文档，规范决策过程

清晰的文档规范有助于系统地执行政策和程序，将机构的偏差管理流程在每个阶段的实施和记录方式标准化，并有助于确保问责制。模型文件应包含对系统机制的可解释性描述，使监督人员能够就系统可能加剧偏差的问题做出明智的、基于风险的决策。文档还可作为重要信息的单一存储库，不仅支持对 AI 系统和相关业务流程的内部监督，还可增强系统维护，并作为任何必要的纠正或调试活动的宝贵资源。

5、完善问责机制，明确权责划分

基本的问责机制即指定某个具体的团队或个人（比如首席模型风险官）来负责 AI 系统中的偏差管控。要求个人或团队为风险及相关危害承担责任，可以直接激励他们去管控风险。问责制需要对 AI 系统本身的作用进行明确评估。例如，决策支持系统一般不负责进行直接决策，因此风险较小，但反而容易被用户过度依赖或被误用、滥用。在这种情况下，AI 系统就会造成与直接参与决策并无二致的危害。此外，模型或算法审计可用于评估和记录这些重要的问责制考虑因素。

6、推广风控文化，探索管控框架

为了使人工智能治理有效，应将治理嵌入到机构的整体文化

当中。风险管理文化与实践可以从社会技术系统的角度来发现整个人工智能生命周期中的偏差。典型的风险管理文化包括：

有效挑战：有效挑战的主要内容是模型风险管理框架的核心组成部分。培育一种有效挑战的文化，鼓励对 AI 系统的开发提出挑战和质疑，有助于在系统部署前就将人工智能偏差问题暴露出来。

三道防线：由于文化很难被直接映射或衡量，鼓励这种方法的一种方式是在机构和程序层面激励批判性思维和审查。模型风险管理框架通常是通过“三道防线”来系统实施的，即建立独立的团队来负责模型生命周期的不同方面。通常第一道防线侧重于模型开发，第二道防线侧重于风险管理，第三道防线侧重于审计，可以帮助机构预测并有效降低偏差风险的出现。

7、优化风险分级，实施激励机制

对人工智能偏差进行有效风险管理的一个核心文化是明确承认风险管理的实质是风险缓解，而非风险规避。培养一种风险缓解意识，即明确接受事故可能发生且将要发生，并在事故发生后强调实际检测和缓解，有助于确保在实践中迅速缓解任何偏差风险。这种认识使得机构能够对风险进行明确分类分级，从而将有限的资源集中在最重要、最可能对现实世界造成损害的偏差风险上。对此，有效的组织文化将其他团队的薪酬和晋升激励措施全

面对标人工智能风险缓解团队，促进风险缓解机制的参与者（比如三道防线）使用合理的开发方法、严格测试和全面审计。

8、推动信息共享，加强经验借鉴

共享网络威胁信息有助于机构改善自身和其他机构的安全态势。制定内部机制，实现偏差事故与有害影响信息共享，有助于提高对人工智能风险的重视程度，提高团队知识和能力，避免过去失败的设计，并防止事故发生。

三、结论

本文广泛论述了人工智能偏差有关的风险，并提出了相应的应对措施。本文基于研究发现，提出如下要点：

1.制定详尽的技术指南需要时间和来自不同相关方的参与，包括与人工智能应用设计、开发和部署相关的群体，也包括可能受到人工智能系统部署影响的群体。

2.采用社会技术视角可以优化人工智能生命周期的流程。人们应将社会价值观付诸实践，并围绕人工智能的构建和部署方式制定出新的规范。为此，需了解计算因素、统计性因素与系统性偏差、人为偏差发生交互作用的机制。

3.为人工智能偏差制定的初步的社会技术框架可按照三个关键领域进行划分和讨论：

（1）数据集在社会技术背景下的可用性、代表性和适用性；

（2）TEVV（测试评估、确认及验证）各环节需要综合考量的注意事项，以及支持测试和评估的相关标准；

（3）人为因素，包括个人和机构内部的社会性偏差和历史性偏差、参与方法（比如以人为本的设计），以及“人在回路”的做法。

译自：*Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, March 2022 by National Institute of Standards and Technology*

译文作者：工业和信息化部赛迪研究院　刘丽超 刘雪宁
联系方式：13466595569
电子邮件：liulc@ccidthinktank.com

赛迪智库
面向政府 服务决策

咨询翘楚在这里汇聚

规划研究所	知识产权研究所	安全产业研究所
工业经济研究所	世界工业研究所	网络安全研究所
电子信息研究所	无线电管理研究所	中小企业研究所
集成电路研究所	信息化与软件产业研究所	节能与环保研究所
产业政策研究所	军民融合研究所	材料工业研究所
科技与标准研究所	政策法规研究所	消费品工业研究所

编 辑 部：工业和信息化部赛迪研究院
通讯地址：北京市海淀区万寿路27号院8号楼12层
邮政编码：100846
联 系 人：王 乐
联系电话：010-68200552 13701083941
传 真：010-68209616
网 址：www.ccidwise.com
电子邮件：wangle@ccidgroup.com

报：部领导

送：部机关各司局，各地方工业和信息化主管部门，
相关部门及研究单位，相关行业协会

编 辑 部：赛迪工业和信息化研究院
通讯地址：北京市海淀区紫竹院路 66 号赛迪大厦 8 层国际合作处
邮政编码：100048
联 系 人：蒯佳佳
联系电话：（010）88559658　18201126359
传　　真：（010）88558833
网　　址：www.ccidgroup.com
电子邮件：kjj@ccidgroup.com