困境工具箱：超越负责任人工智能/机器学习中的去偏与公平性公式

Andrés Domínguez Hernández布里斯托大学计算机科学系，英国布里斯托
andres.dominguez@bristol.ac.uk

Vassilis Galanos爱丁堡大学科学、技术与创新研究系英国爱丁堡
vassilis.galanos@ed.ac.uk

摘要——**近年来，对公平与合乎伦理的人工智能的研究方法受到了新兴的、以定性研究为主的批判性数据研究领域的审视；该领域强调此类干预缺乏对情境和复杂社会现象的敏感性。我们借鉴其中一些启示，提出一个由三部分组成的决策工具箱，其内容来自追求负责任的人工智能/机器学习时遇到的困境，具体包括：(a) 机会困境：数据的可用性决定问题陈述，还是问题陈述决定数据收集与处理；(b) 规模困境：可扩展性与情境化能力之间的取舍；以及 (c) 认识论困境：实用主义技术客观主义与反思性相对主义之间的张力，关乎如何承认社会因素。本文主张，应基于具体情境进行推理，并创造性地应对负责任的算法／数据驱动系统所面临的困境，超越公平人工智能文献中常见的公式化偏见消除和伦理操作化叙事。**

关键词——数据标注、伦理、公平性、负责任的人工智能、科学技术研究

I. 引言NTRODUCTION

A. 研究缘由、背景与提纲

大量近期的批判性研究表明，基于有偏数据集、存疑的科学依据训练且未考虑情境的机器学习（ML）算法系统，会对不同人口群体产生不均衡的影响，并对边缘化群体造成不成比例的伤害 [1]–[3]。计算机科学和数据科学领域的研究人员针对伦理关切以及社会科学的批评，提出了公平性、风险评分和可解释性框架 [4]–[7]。这类研究主要聚焦于开发公平或无偏的模型、将伦理框架付诸实践，以及改进文档和可理解性。然而，由于对现实条件的评估不足，或开发者不熟悉算法产生影响的情境，这种以计算为导向的方法可能会陷入公平性定义相互冲突、以及对何为或应为何种道德上正确的结果意见不一的困境 [8], [9]。人工智能和机器学习开发仍未充分反思支撑算法系统开发的认识论、动机和政治因素如何影响其社会影响，尤其是在人工智能开发的优先事项设定方面 [10]。我们认为，现有的负责任人工智能/机器学习框架仍需处理一系列问题，而知识生产领域的既有理论和研究，尤其是科学技术研究（STS），早已探讨过这些问题。本文深入探讨算法系统诸多应用领域中固有的矛盾、模糊性和不确定性 [9], ，并希望提供一套工具，助力负责任的决策。我们提出，应将棘手挑战视为契机，通过立足具体情境、富有创意地审视困境，协商达成合乎伦理的结果。我们提出，在追求负责任地开发算法系统时会面临三种难题，并将在本文第二节进一步分析和剖析：（a）机会困境，即数据集和问题陈述，究竟孰先孰后？（b）规模 困境，即扩大规模的成本效益压力与依据社会特殊性进行情境化、具体化的要求之间的张力；以及（c）认识论困境，即如何在朴素客观主义测量与不切实际的相对主义评估之间确定负责任的干预方式。我们的主张源于长期与人工智能/机器学习从业者合作的经历；作为协作研究项目的成员，我们也借鉴了科学技术研究及相关领域的批判性学术研究成果。我们提出一个理论框架，建立在人工智能/机器学习设计中不同实践情境所共有的主题之上。我们认为，在围绕人工智能/机器学习应用的设计局限展开讨论和撰写报告时，应将该框架中的困境纳入其中，从而推动过程更加透明、负责任。我们希望借此倡议，促成围绕数据驱动技术当下的合法性危机、真相与信任问题，以及更广泛而言算法系统的认识论基础和炒作性主张，展开批判性思考与跨学科对话。

II. THEORETICAL 贡献: THE THREEDILEMMAS 适用于人工智能 RESEARCHERS

本节概述当代计算科学与社会科学中有关负责任创新的讨论所引出的三种困境。我们认为，尽管这些困境看似棘手且无法解决，但不应将其视为障碍，而应视为人工智能/机器学习从业者需要留意的参数。同样，我们指出，这些困境并非面面俱到，也不会在所有情境中以同等方式出现：某些领域的解决方案可能比其他领域更直接。在负责任创新实践中，一种可能的应用方式[11]–[13] 是审查从业者或设计人员的回应，以及他们选择某个选项而非另一选项的合理化论证；或者，在适用的情况下，引入创新方法（包括非技术方法）来化解困境。表1总结了这三种

种拟议困境及其选项，并简要概述各自的优缺点。不应将此类困境视为泾渭分明的非此即彼选择，而应将其视为成对对立关系的一部分，由此引出新的知识表征。正如近期在通过辩证综合将知识管理与数据标注实践相结合的讨论中所提出的那样，两种表面上的矛盾可以成为社会变革的契机[11]。

A. 机会困境：什么应当优先？数据还是问题陈述？

我们识别出的第一个困境涉及人工智能开发中的创新与优先事项由什么驱动。问题陈述是否会塑造所需数据集的定义与准备？还是说，数据的机会性可用性以及对现有数据可供性的考察，会帮助塑造人工智能要解决的问题？人工智能领域的大量研究受到技术解决主义的驱动，这种解决主义以技术上、经济上乃至政治上可行的方案为基础。数据可用性是开发应用的关键标准。如果数据可用性先于人工智能待解决问题的提出，那么这或许主要是以机会为导向的；由现有数据造成的约束与路径依赖所引发的问题化，便会沦为次要考量。比如，考虑一下使用基于二元性别分类的遗留数据库，来自动化关于跨性别者和非二元性别者权利的决策，会产生什么影响——无论是依据人脸识别数据集 [12], 还是身体扫描X射线数据集 [13]来进行性别／性别认同识别，都存在这一问题。这样的数据优先方法可能会导致有害的不匹配，即身体特征与通常和特定性别认同相关联的特征不相符（此处可考虑乳腺癌机器学习筛查的影响 [14]）。另一种做法是从理解社会群体的需求与脆弱性出发；在此例中，该群体是LGBTQI+ 社群。这或许需要构建能够反映这类多样性的新数据集——这种做法需要投入更多资源，并通过政治游说争取这些资源。面对这一困境，我们应当追问：正在开发的应用是否只回应机会逻辑与技术推动，是否旨在满足迫切的社会需求，抑或两者兼顾？此外，这一困境还要求我们思考数据集标注与聚类过程中常规性和偶然性的问题。正如定性研究所表明的，日常标注工作鲜少是由了解数据集合之社会意义且具备先验知识的技术专家所开展的信息充分的实践 [15], [16]。相反，标注往往是脆弱而平凡的过程，例如，工作人员投入更多时间创建聚类数据的理想类型，却较少反思某一标签是否具有表征真实性。数据工作很可能是在赶工期的压力下进行的，其结果也容易受到数据科学家对特定聚类象征价值的理解所影响，因而数值输出具有解释灵活性。换言之，并非每位数据科学家都能了解特定聚类或数据集的社会语境，尤其是在某个聚类的数值表征使标注者与数据收集时所处的情境相疏离。

B. 规模困境：泛化还是情境化？

在关于负责任的人工智能/机器学习的讨论中，关于数据存在两种相互竞争的观点：越多越好，以及少即是多。一些公平性的定义认为，更多且质量更高的数据将减轻算法偏见与不公正。这正是传统数据科学所偏好的方法：它旨在满足降低伤害的成本与道德要求，同时通过更大规模的数据获得更准确的发现，从而提高经济回报。然而，越多越好的论点较少关注“数据过多”可能对参与标注工作的人造成有害影响的问题，正如若干批判性案例研究所表明的那样 [1], [2]。数据过多在技术上也可能产生不利影响。例如，在机器学习的医学影像应用中，近期一项研究指出，“更多样本可能导致更差的性能”，因为它们会掩盖细微却重要的结构变化、不稳定性以及基于各种社会因素产生的几乎难以察觉的扰动；这些现象在大规模模式分析中很容易被忽视，因此作者呼吁对这类不稳定现象开展更多基于情境的测试 [17]。另一种方法则受到社会要求的影响，主张将数据集置于其既定情境中，以满足受算法系统影响的特定群体和代表性不足群体的需求与要求 [18]。后者的少即是多 方法认为，有情境依托的小规模定性方法和参与式方法是必要的，非技术性解决方案也应成为负责任创新的一部分。然而，通过更具情境依托的方法缩小规模并优先考虑情境准确性，可能面临资金障碍，存在参与洗白或给边缘化群体带来认知负担的风险，也可能难以扩大规模 [19], [20]。此外，过度具体的数据集违背了建模的根本目的，破坏了地图与疆域关系；因此，人们期待基于情境的数据集和模型至少具备一定程度的可泛化性。在机器学习领域，至少自1980年以来，人们就已知晓需要综合运用“先验知识、偏见和观察来指导学习过程”[21], ，而这一点又反馈到第一个困境的进程中。在思考这一困境时，研究人员需要重新审视那种普遍存在且被视为理所当然的扩大规模信条，并考虑针对特定目的、在数据从业者、数据生成者和数据所有者之间实现更广泛社会协同的替代方案 [22] 。

C. 认识论困境：客观主义还是相对主义？

我们所确定的第三个困境，存在于这样一种更深层的认识论立场之中：设计者或数据科学家认同一种负责任的干预方法，将算法设计建立在成本效益、性能和客观性等实证主义、技术中心主义价值观之上 [23]。数据科学所蕴含的实用主义表明，这种方法是适当的，并且在很大程度上具有客观性，至少只要存在能够为数据集的真实性提供合理辩护的方式。另一方面，社会科学批判则会将客观性问题化

表 *1*：困境工具箱

困境工具箱		
机会困境	规模困境	认识论困境
数据集之后问题陈述	可泛化数据集	客观主义者测量
优势： 稳定的决策 使实践成为可能 缺点： 持续更新的 数据和更广泛的 语境/语义， 路径依赖 静态遗留 数据集的影响 问题化 并延续 不公正现象	优势：成本——效率。改进后的推荐 基于多样性 缺点： 比例失衡的 代表性 （以及对.....的伤害） 边缘化群体 以及特定情境 需求；可解释性 方面的考量、易于 使用数据集 朝着理想的 结果。不 是所有事情都应该被 量化	优势：极高的 功利主义精确度 指标 缺点： “电脑说了算” 效应； 施为性 误导性指标 性能； 对.....造成伤害 被边缘化的， 未被代表的 社群
问题陈述随后是数据集	特定情境数据集	相对主义批判
优势： 信息充分 负责任的 方法， 与正当的 社会需求相符的 应用， 技术 这种方法并非 先决条件 缺点： 可能缺乏 合适的数据集， 经济 不可行性	优势： 对 需求特殊性 以及.....的特殊性 社会语义 劣势： 资金障碍， 认知负担， 缺乏可泛化 效用	优势： 承认 社会丰富性， 权力和 对.....的意识 多重挑战。 标记有害的 概括 缺点： 高层次、宽泛 且含糊的 推荐。 不明确 适用性。 实践上的徒劳 高速技术 部署

，将其视为一种天真的理想，并强调任何科学事业都受社会、文化、政治和物质因素制约的特性。然而，这种批判很容易陷入无休止的争论和“智力游戏”之中，而这常常被认为会阻碍实现具有社会积极意义的结果或交付产品的实用需求 [24]–[26]。正如先前从“科学战争”争论中得到的教训所示，客观主义科学已被证明并非独立于社会语境的权威，而相对主义科学则有可能削弱付诸实践的行动号召，并被反科学行动者加以利用 [27]。对于实用却天真的客观主义与反思却徒劳的相对主义之间的张力，可以通过批判性的学科重构找到出路；这些重构面向技术—社会变革，例如 Birhane 提议将计算科学中的理性形式主义转向关系伦理方法 [28], ，或日益高涨的呼声——将反思性和定性方法引入算法系统的设计之中 [29]–[31]。

III. Conclusions 和 Further Work

过去十年间，社会科学各领域涌现出大量深入且细致的研究，探讨用于构建真实值数据集的“真相”来源是否可争议、算法政治、策展过程所隐含的无形物质与人力成本、算法指标的施为性，以及人工智能未来想象，以及对个人/机构动机的评估，考察其与社会和/或特定用户利益的关系 [9], [30], [32]。然而，要将这些丰富的社会科学研究融入人工智能/机器学习开发、监管和教育，仍存在巨大鸿沟。为助力应对这一挑战，我们在此呈现了追求负责任、公平且公正地开发算法系统时所产生的一系列困境。尽管困境这一框架通常会让人感到分析陷入瘫痪，或是通过缺乏充分论证的平衡与取舍公式迅速加以化解，但我们建议，在讨论负责任的人工智能/机器学习时，应对这些困境进行深入而持续的审视。禅宗逻辑通过向参禅者提出无解的悖论谜题（公案），为他们创造了克服悖论、进而获得开悟的机会。将日常困境视为冥想悖论的做法，已在临床实践中得到有效运用 [33] ——为何不也应用于机器学习实践？设计者希望依据现有数据集来定义问题，还是寻找符合现有问题定义的数据集？我们想要可泛化、成本有效的数据，还是专业化、针对特定情境的数据？我们更倾向于客观主义的天真，还是相对主义的徒劳？如果将这些问题视为可以相互权衡的机会，从而探索细致入微的新选择，并以此实现类似辩证综合的效果 [11]，或许最佳答案并不局限于简单的二元对立。我们希望，这些审视能够激发富有创造性且意义深远的讨论，并避免预设结论，以及盲目接受人工智能/机器学习中占主导地位的、宣扬必然性与解决主义的话语。

Acknowledgment

我们声明，本文撰写未获得任何资金支持，且在此过程中未发现任何利益冲突。ADH 谨感谢 REPHRAIN 在 UKRI 项目资助 EP/V011189/1 下，于本文撰写期间提供的支持。

References

[1] E. M. Bender、T. Gebru、A. McMillan-Major 和 S. Shmitchell，“论随机鹦鹉的危险：语言模型会不会太大？🦜”，载于*2021*年*ACM*公平、问责与透明度会议论文集，美国纽约州纽约，2021年3月，第610–623页。doi: 10.1145/3442188.3445922。[2] V. U. Prabhu 和 A. Birhane，“大型图像数据集：计算机视觉的一场得不偿失的胜利？”，arXiv，2020年7月23日。访问日期：27, 2022 [ ] 2006 16923 年5月。[3]
在线 . 获取地址：http://arxiv.org/abs/ 。L. Stark 和 J. Hutson，“相面人工智能”，社会科学研究网络，纽约州罗切斯特，SSRN学术论文编号3927300，2021年9月。doi: 10.2139/ssrn.3927300。[4] M. Feldman、S. A. Friedler、J. Moeller、C. Scheidegger 和 S. Venkatasubramanian，“认证并消除差别影响”，载于第*21*届*ACM SIGKDD*国际知识发现与数据挖掘会议论文集，美国纽约州纽约，2015年8月，第259–268页。doi: 10.1145/2783258.2783311。[5] M. Geisslinger、F. Poszler、J. Betz、C. Lütge 和 M. Lienkamp，“自动驾驶伦理：从电车难题到风险伦理”，《哲学与技术》，第34卷，第4期，第1033–1055页，2021年12月，doi: 10.1007/s13347-021-00449-4。[6] M. Hardt、E. Price 和 N. Srebro，“监督学习中的机会平等”，载于第*30*届国际

神经信息处理系统会议，美国纽约州雷德胡克，2016年12月，第3323–3331页。[7] F. Kamiran 和 T. Calders，“无歧视分类”，载于控制与通信*2009*年第二届计算机国际会议，2009年2月，第1–6页。doi: 10.1109/IC4.2009.4909197。[8] A. D. Selbst、D. Boyd、S. A. Friedler、S. Venkatasubramanian 和 J. Vertesi，“社会技术系统中的公平与抽象”，载于公平、问责与透明度会议论文集，美国纽约州纽约，2019年1月，第59–68页。doi: 10.1145/3287560.3287598。[9] A. Birhane，“歧义自动化的不可能性”，人工生命，第27卷，第1期，第44–61页，2021年6月，doi: 10.1162/artl a 00336。[10] I. Lowrie，“算法理性：数据科学中的认识论与效率”，《大数据与社会》，第4卷，第1期，第2053951717700925页，2017年6月，doi: 10.1177/2053951717700925。[11] E. Rusyaeva 和 A. G. Kravets，“用于知识管理的创造性知识表示：辩证方法”，载于智能技术与数据科学中的创造力，查姆，2021年，第97–109页。doi: 10.1007/978-3-030- 87034-8 8。[12] D. Anderson，“GLAAD和HRC呼吁斯坦福大学及负责任的媒体揭穿一份危险且有缺陷的报告，该报告声称可通过面部识别技术识别LGBTQ人群”，*GLAAD*，2017年9月08日。https://www.glaad.org/blog/glaad-and-hrc-call-stanford-university-responsible-media-debunk-dangerous-flawed-report（访问日期：2022年5月30日）。[13] S. Costanza-Chock，《设计正义：构建我们所需世界的社区主导实践》。麻省理工学院出版社，2020年。访问日期：2022年3月02日。[在线]。获取地址：https://library.oapen.org/handle/20 500 12657 43542 。. / E. D. [14] Pisano，“人工智能在乳腺癌筛查中展现前景”，《自然》，第577卷，第7788期，第35–36页，2020年1月，doi: 10.1038/d41586-019-03822-8。[15] D. M. Kotliar，“社会的回归：象征性消亡时代的算法身份”，《新媒体与社会》，第22卷，第7期，第1152–1167页，2020年7月，doi: 10.1177/1461444820912535。[16] M. Miceli 和 J. Posada，“数据生产装置”，《*ACM*人机交互汇刊》，第6卷，第CSCW2期，2022年。[17] V. Antun、F. Renna、C. Poon、B. Adcock 和 A. C. Hansen，“图像重建中深度学习的不稳定性及人工智能的潜在成本”，《美国国家科学院院刊》，第117卷，第48期，第30088–30095页，2020年12月，doi: 10.1073/pnas.1907377117。[18] M. Katell 等人，“迈向算法公平的情境化干预：来自实践领域的经验教训”，载于*2020*年公平、问责与透明度会议论文集，美国纽约州纽约，2020年1月，第45–55页。doi: 10.1145/3351095.3372874。[19] J. Pierre、R. Crooks、M. Currie、B. Paris 和 I. Pasquetto，“共同协作：以数据为中心的参与式设计研究与认知负担”，载于*2021*年*CHI*会议论文集计算系统中的人为因素会议，美国纽约州纽约，2021年5月，第1–11页。doi: 10.1145/3411764.3445103. [20] M. Sloane、E. Moss、O. Awomolo 和 L. Forlano，《参与并非机器学习的设计补救措施》，arXiv，arXiv:2007.02423，2020年8月。doi: 10.48550/arXiv.2007.02423. [21] T. Mitchell，《学习概括需要偏见》，2002年10月。[22] B. Ribeiro 等，《提出包容且负责任的研究与创新中社会对齐的困境》，*null*，第5卷，第3期，第316–331页，2018年9月，doi: 10.1080/23299460.2018.1495033.[23] A. Birhane、P. Kalluri、D. Card、W. Agnew、R. Dotan 和 M. Bao，《机器学习研究中编码的价值观》，载于*2022*年*ACM*公平、问责与透明度会议，美国纽约州纽约，2022年6月，第173–184页。doi: 10.1145/3531146.3533083.[24] D. Moats 和 N. Seaver，《“你们社会科学家就爱玩心理游戏”：在数据科学与批判性算法研究之间的“鸿沟”中进行实验》，《大数据与社会》，第6卷，第1期，文章编号2053951719833404，2019年1月，doi: 10.1177/2053951719833404. [25] D. Collingridge 和 C. Reeve，《科学对权力发声：专家在政策制定中的作用》，1份在线资源（xi，175页）。纽约：圣马丁出版社，1986年。访问日期：2022年5月30日。[在线]。可访问：https://archive.org/details/sciencespeakstop0000coll [26] S. Andreski，《社会科学作为巫术》。纽约：圣马丁出版社，1973年。访问日期：2022年5月19日。 [在线]。可访问：http://archive.org/details/socialsciencesas00andr [27] M. Lynch，《我们从未反科学：对科学战争与后真相的思考》，参与科学技术与社会研究，第6卷，第49–57页，2020年1月，doi: 10.17351/ests2020.309. [28] A. Birhane，《算法不公：一种关系伦理方法》，*Patterns*，第2卷，第2期，文章编号100205，2021年2月，doi: 10.1016/j.patter.2021.100205. [29] C. D’Ignazio 和 L. F. Klein，《数据女性主义》。麻省理工学院出版社，2020年。[30] M. Miceli、T. Yang、L. Naudts、M. Schuessler、D. Serbanescu 和 A. Hanna，《记录计算机视觉数据集：对反思性数据实践的倡议》，载于*2021*年*ACM*公平、问责与透明度会议论文集，美国纽约州纽约，2021年3月，第161–172页。doi: 10.1145/3442188.3445880.[31] A. Tanweer、E. K. Gade、P. M. Krafft 和 S. K. Dreier，《数据革命为何需要定性思维》，《哈佛数据科学评论》，第3卷，第3期，2021年7月，doi: 10.1162/99608f92.eee0b0da.[32] A. Kerr、M. Barry 和 J. D. Kelleher，《对人工智能的期待与伦理的施为性：对传播治理的影响》，《大数据与社会》，第7卷，第1期，文章编号2053951720915939，2020年1月，doi: 10.1177/2053951720915939. [33] C. A. Simpkins 和 A. M. Simpkins，《心理治疗中的禅修：临床实践技巧》。新泽西州霍博肯：约翰·威 Wiley父子公司，2012年。访问日期：2022年5月30日。[电子资源]。可访问：http://ezprox.bard.edu/login?url=http://ebookcentral.proquest.com/lib/connectny/detail.action?docID=817380