迈向女性主义交叉性可解释人工智能：从可解释性到回应能力

Goda Klumbytė
卡塞尔大学电气工程与计算机科学学院，
德国 goda.klumbyte@uni-kassel.de

Hannah Piehl
卡塞尔大学电气工程与计算机科学学院，
德国
hannah.piehl@stud.uni-frankfurt.de

Claude Draude
德国卡塞尔大学电气工程与计算机科学学院 claude.draude@uni-kassel.de

允许免费为个人或课堂使用而制作本作品部分或全部内容的数字副本或纸质副本，前提是不得为营利或商业利益而制作或分发副本，且副本首页须载有本声明和完整引文。本作品中第三方组件的版权必须得到尊重。其他所有用途，请联系版权所有者/作者。

Goda Klumbytė、*Hannah Piehl* 和 *Claude Draude*。*2023*年。《迈向女性主义交叉性可解释人工智能：从可解释性到回应能力》。工作坊：以人为中心的可解释人工智能（*HCXAI*），计算系统中的人为因素会议 *CHI '23*，*2023*年*4*月*23*日至*28*日，德国汉堡，*9*页。© 2023 版权所有归所有者/作者所有。

摘要
本文响应了关于采用计算领域的批判性方法，以及对交叉性、女性主义、去殖民化人机交互和人工智能设计进行概念化的呼吁，并探讨了在 HCXAI 研究与设计中，女性主义交叉性视角可能呈现为何种面貌。本文勾勒出初步的研究方向及其对可解释人工智能设计的启示，指出从女性主义视角来看，可解释性应包括培养回应能力——即批判性评估人工智能系统并对此作出回应的能力——同时应将边缘化视角置于核心位置。

作者关键词
女性主义交叉性；以人为中心的可解释人工智能；女性主义可解释人工智能。

CSS 概念
• 以人为中心的计算 • 人机交互（HCI）● 人机交互理论、概念与模型

引言
过去十年间，人机交互和人工智能研究见证了批判性跨学科研究方法的兴起：人们纷纷呼吁开展女性主义交叉性人机交互 [4, 5, 7, 34, 44] 与人工智能/机器学习研究 [11, 31, 49, 51], 后殖民与去殖民人机交互研究 [6,

29, 42] 和人工智能 [1, 37], 以残障为中心的人工智能 [18, 51],人机交互中的批判种族视角 [41, 45] 和人工智能 [20,35]。本文通过探讨可解释人工智能（XAI）中的女性主义交叉性方法可能呈现何种形式，为这些倡议作出了贡献。

情境化知识与立场理论

倡导以交叉性的视角理解社会定位和结构性不平等。边缘化的观点能够为压迫、权力和特权的运作方式提供特定视角，也可以成为另类知识实践的立足点。立场理论还促使研究者追问：特定研究是从谁的视角出发、为谁的利益而开展的，并要求优先考虑边缘化群体的需求和利益。

我们认为，可解释人工智能中的女性主义交叉性视角与以人为中心的可解释人工智能（HCXAI）相契合：后者关注价值观、人际动态、人工智能的社会情境性，并致力于建立一种反思性的社会技术实践 [12, 13, 14, 15, 16]。这一视角进一步提出了可解释人工智能的以社会正义为中心的规范取向，要求关注权力差异和结构性不平等，挑战可解释人工智能的理性主义普遍主义模式，并培养用户批判性评估人工智能系统并作出回应的能力，即回应能力。这一概念视角也对可解释人工智能设计具有实际意义，因为它将用户多样性和可解释人工智能生态系统中的权力动态纳入考量。

HCXAI 中**社会技术转向**的情境取向——鼓励密切关注人工智能的组织与社会技术背景，以支持构建可解释人工智能系统，并为用户的批判性解读与反思留出空间——与女性主义认识论立场相契合。

女性主义认识论与（以人为中心的）可解释人工智能中的女性主义视角

女性主义认识论植根于情境化知识和立场理论，这些理论承认，（科学）知识总是置身于特定的物质、学科、具身、文化、社会政治和历史语境之中 [21]。因此，知识始终是局部的。局部的情境化知识反过来提供了一种更立足现实的真理观，并使问责成为可能，因为能够定位的知识主张，也能够向其影响或涉及的人们和语境负责。因此，情境知识不再追求价值中立、普世的客观性，而是提供“作为立场化理性的客观性”[21, p.590] 以及多元的立足现实的视角。

女性主义立场理论认为，知识不仅是在特定社会定位中生产的，而且深受价值观影响。它们提出，要迈向公平与正义，知识生产工作应优先考虑边缘化立场 [22, 23, 24, 25, 26, 27]。立场或社会定位并非单一维度，因为类别归属的多个轴线（性别、种族、阶级、性取向、残障／能力等）以及由此产生的特权或压迫会在这些定位中交织。这会产生特定形式的压迫与不平等，而交织的压迫会带来特定形式的歧视 [8], ，因此需要对公平作交叉性理解 [17] ，并认识到用户占据着多维身份的交汇位置。

先前研究记录了 HCXAI 中的社会技术转向，并提出了一项议程，主张关注人工智能系统的价值观、人际动态和社会情境性，以此推动形成反思性的社会技术实践 [14]。研究者指出，可解释人工智能具有情境依赖性，因为可解释人工智能面向不同的用户群体——例如新手用户、数据专家和人工智能专家 [38] ——他们对于何为良好解释有着不同的解释目标和评价标准 [39, 40, 50]。这与交叉性的概念相呼应：交叉性建立在这样一种观点之上，即身份和社会定位既多元又相互交织；情境知识和立场理论的视角则强调知识主张的嵌入性和情境依赖性。Singh 等人 [46] 主张突出全球南方的可解释人工智能视角，挑战白人西方理性主义解释框架，并提出可解释性会呈现何种面貌

关注权力与结构性不平等

这意味着要提出诸如此类的问题：

- 这些解释是为谁的利益而生成的，又是从谁的视角出发的？
- 它们将对谁开放并发挥作用？
- 还有哪些内容未得到解释？

在可解释人工智能的语境下，这些问题首先应由设计人员提出，但对所有利益相关者都有益。

面向印度农村地区边缘化群体。女性主义交叉性视角倡导审视并挑战知识生产中的权力等级，这一主张也与这一立场相契合。

有少数论文主张在可解释人工智能中明确采用女性主义视角[19, 28, 32, 33]。Hancox-Li 等人 [19]探究了特征重要性方法中的认识论价值，指出女性主义认识论原则中的多元性、关注社会背景以及优先考虑受压迫群体视角，可以引导可解释人工智能纳入受压迫群体的视角。Huang 等人 [28] 也主张关注情境，强调可解释人工智能设计需要跨学科性，以丰富用于构建更好解释和识别偏见的"诠释资源"与背景假设，并呼吁更多不同利益相关者群体参与。Singh 等人[47] 提及交叉性，并采用更加以用户为中心的方法，聚焦情境，同时考量解释框架及其适切性。

迈向女性主义交叉性可解释人工智能

我们通过概述一系列可解释人工智能的视角及其对可解释人工智能设计的后续影响，勾勒出女性主义交叉性可解释人工智能的雏形。我们将这一雏形作为一种总体方法，旨在从方向上为可解释人工智能研究与设计提供概念、理论和规范层面的根基，使其朝向社会正义。

可解释人工智能的女性主义原则

以社会正义和公平为规范性导向。女性主义认识论认为，知识、技术设计和技术实践并非价值中立，而是社会技术现实的产物，反映了特定的价值观和议程。

女性主义可解释人工智能将突出有助于实现正义与公平的可解释人工智能设计，使可解释人工智能与公平、问责、透明度和伦理（FATE）密不可分。女性主义立场指出，仅仅关注不平等并不足够；人工智能系统和解释工作应致力于积极消除不平等，并推动构建正义、公平的社会技术人工智能生态系统。

关注权力与结构性不平等。可解释人工智能研究与设计还应研究、理解并考量：在特定的社会技术情境中，与可解释性和解释设计相关的概念如何与权力动态交织。这种交织可能以不同方式显现：再现权力动态和既有不平等，引入新的社会紧张关系与不平等，或主动削弱这些现象。

挑战普遍主义和传统理性主义解释模式。女性主义认识论和交叉性支持视角的多元性，并摒弃普遍理性的观念，因此，仅专注于生成关于人工智能系统的认知性解释，并不足够。相反，正如黑人女性主义、后殖民、去殖民和原住民视角所表明的，西方理性形式具有文化特殊性。因此，女性主义可解释人工智能方法应致力于拓展解释的含义，以及解释的呈现方式。

以边缘化视角为中心。 在可解释人工智能中，寻找以边缘化视角为中心的方法至关重要，

理性主义解释模态 以理性作为解释基础的理解为依托。这些模态意味着理性具有普遍性，因此不同情境中的不同人群都具有相同的理性。普遍理性的观念受到批评，因为它对人格的理解过于简化，也与通过将某些群体称为“较不理性”而对其进行非人化的做法相容（历史上，这些群体包括女性、被种族化的人群、LGBTQ群体及其他边缘化群体）[36]。

回应能力 entails
探讨可解释人工智能研究人员和设计人员应向哪些社群、哪些用户群体、哪些身体负责，以及如何设计可解释人工智能，从而为这些社群和群体创造更多可能性，使其能够在与人工智能的交互中进行批判性且富有建设性的回应 [10]。

也就是说，通过鼓励可解释人工智能设计中的参与式和包容性流程，促进跨学科性，并积极推动开发有益于边缘化利益相关者、满足其需求的可解释人工智能。此外，这还要求生成能够积极赋能用户的可解释人工智能系统，使用户能够批判性地解读、评估，并且尤为重要的是，对人工智能系统作出回应。这种回应能力正是女性主义情境知识理论所提出的评估问责的关键框架 [2, 3, 21]。

对可解释人工智能设计的启示

- 可解释人工智能设计人员可以对可解释人工智能设计的目标、手段和目的采取情境导向的探究，特别关注特定人工智能应用领域和/或情境中的权力动态与结构性不平等。

- 这要求设计人员关注系统层面，即不仅要解释模型本身，还要理解其情境以及特定人工智能系统背后的更广泛依据。决定某一特定人工智能系统及其影响从何处开始、到何处结束的边界工作，同样需要这种系统性视角；团队拥有更大的学科多样性有助于实现这一点。

- 在整个（以及超越）所有设计阶段的女性主义可解释人工智能设计过程中，从一开始就注重主动性以及交叉性视角至关重要 [9, 42, 49]。

- 参与式和交互式解释设计对于理解并考量边缘化视角、培养批判性回应能力至关重要；同时，还应推动可解释人工智能设计团队多元化，并致力于运用跨学科视角（例如社会科学和人文学科的知识）来制定设计目标和评估工具，以更好地支持将边缘化处境者的需求付诸实践。

- 通过整合交叉性来理解不同的利益相关者和用户群体，也可以借助更多参与式设计方法加以促进，从而践行以受压制的知识为中心的女性主义原则。

- 拓展解释模态意味着探索超越更传统的理性认知型解释的解释传达方式，这也是一项重要启示。

讨论

最后，我们提出一些问题，以供讨论和进一步研究。尽管女性主义交叉性视角能为可解释人工智能提供更广泛的指导，但它在多大程度上能够被形式化为具体的应用方法？即便如此，这些方法又能在多大程度上推广到不同情境中？我们认为，女性主义交叉性可解释人工智能需要抵制可扩展性，转而聚焦于个案特异性 [30, 48]。女性主义交叉性可解释人工智能在多大程度上能够应用于商业化程度更高、以利润为导向的可解释人工智能系统？由于女性主义交叉性视角本质上具有政治性且以规范为导向，更广泛的采纳是否可能？这需要什么条件？可解释人工智能中的女性主义交叉性视角，能否也与去殖民、后殖民及其他非欧洲中心主义视角进行富有成效的对话？抑或它仍受限于其不可避免地植根于特定（全球北方）情境这一事实？我们认为，女性主义交叉性可解释人工智能是一个有待进一步探索、也亟需深入研究的领域。

致谢

本研究得到大众基金会“人工智能与未来社会”资助，属于合作项目“人工智能取证：通过可解释性实现视觉人工智能系统的问责”的一部分。

参考文献

[1] Syed M. Ali。2016年。《去殖民化计算简论》。*XRDS* 22，4，16–21。数字对象标识符：https://doi.org/10.1145/2930886。[2] Karen Barad。2007年。《与宇宙相遇：量子物理学与物质和意义的纠缠》。杜克大学出版社，达勒姆。[3] Karen Barad。2012年。“内在作用”（Adam Kleinmann 对 Karen Barad 的访谈）。*Mousse*，76–81。[4] Shaowen Bardzell。2010年。《女性主义人机交互：总结现状并为设计勾勒议程》。载于*CHI2010*。计算系统中的人为因素会议第*28*届年会论文集：论文与笔记。ACM，纽约，纽约州，1301–1310。数字对象标识符：https://doi.org/10.1145/1753326.1753521。[5] Shaowen Bardzell 和 Jeffrey Bardzell。2011年。《迈向女性主义人机交互方法论：社会科学、女性主义与人机交互》。载于计算机系统中的人为因素*SIGCHI*会议论文集。ACM，纽约，纽约州，675–684。数字对象标识符：https://doi.org/10.1145/1978942.1979041。[6] Paula Chakravartty 和 Mara Mills。2018年。《“去殖民化计算”虚拟圆桌讨论》。*Catalyst* 4，2，1 4–。数字对象标识符：https://doi.org/10.28968/cftt.v4i2.29588。[7] Sarah Ciston。2019年。《交叉性人工智能至关重要：多声部、多模态、实验方法，以拯救人工智能》。*CITAR*期刊 11，2，3–8。数字对象标识符：https://doi.org/10.7559/citarj.v11i2.665。[8] Kimberle Crenshaw。1989年。《消除种族与性别交叉性的边缘化：对反歧视原则、女性主义理论与反种族主义政治的黑人女性主义批判》。芝加哥大学法律论坛 1989年，8。[9] Mathieu d'Aquin、Pinelopi Troullinou、Noel E. O'Connor、Aindrias Cullen、Gráinne Faller 和 Louise Holden。2018年。面向人工智能研究项目的“设计即伦理”方法论。载于*2018*年*AAAI/ACM*人工智能、伦理与社会会议论文集。ACM，[S.l.]，54–59。数字对象标识符：https://doi.org/10.1145/3278721.3278765。[10] Claude Draude。2020年。“边界并非一成不变”：从交互到人机交互中的能动性内在作用。载于《人机交互：设计与用户体验》。HCI 2020，作为第22届国际会议的一部分举行，论文集，第一部分。信息系统与应用（含互联网／万维网）及人机交互，12181。Springer，20–32。数字对象标识符：https://doi.org/10.1007/978-3-030-49059- 1 2。_[11] Claude Draude、Goda Klumbytė、Phillip Lücking 和 Pat Treusch。2019年。情境化算法：一种应对偏见的社会技术系统方法。《在线信息评论》 44，2，325–342。数字对象标识符：https://doi.org/10.1108/OIR-10-2018-0332。

[12] Upol Ehsan、Q. V. Liao、Michael Muller、Mark O. Riedl 和 Justin D. Weisz。2021。《扩展可解释性：迈向人工智能系统的社会透明性》。载于 *CHI'21*。《2021年CHI计算系统中的人为因素会议论文集：乘势而上：2021年5月8日至13日，线上虚拟会议（原定于日本横滨举行）。计算机协会，纽约，纽约，1–19。数字对象标识符：https://doi.org/10.1145/3411764.3445188。

[13] Upol Ehsan、Samir Passi、Q. V. Liao、Larry Chan、I-Hsiang Lee、Michael Muller 和 Mark O. Riedl。2021。《可解释人工智能中的“谁”：人工智能背景如何塑造人们对人工智能解释的看法》。[14] Upol Ehsan 和 Mark O. Riedl。2020。《以人为中心的可解释人工智能：迈向反思性的社会技术方法》。载于 *HCI International 2020——*最新研究论文：多模态与智能。Springer，Cham，449–466。数字对象标识符：https://doi.org/10.1007/978-3-030-60117-1_33.[15] Upol Ehsan、Philipp Wintersberger、Q. V. Liao、Martina Mara、Marc Streit、Sandra Wachter、Andreas Riener 和 Mark O. Riedl。2021。《将以人为中心的视角付诸实践：可解释人工智能》。载于 *2021*年*CHI*计算系统中的人为因素会议扩展摘要。ACM数字图书馆。计算机协会，纽约，纽约，美国，1–6。数字对象标识符：https://doi.org/10.1145/3411763.3441342。[16] Upol Ehsan、Philipp Wintersberger、Q. V. Liao、Elizabeth A. Watkins、Carina Manger、Hal Daumé III、Andreas Riener 和 Mark O. Riedl。2022。《以人为中心的可解释人工智能（HCXAI）：超越打开人工智能黑箱》。载于 *CHI*计算系统中的人为因素会议扩展摘要。ACM数字图书馆。计算机协会，纽约，纽约，美国，1–7。数字对象标识符：https://doi.org/10.1145/3491101.3503727。[17] James R. Foulds、Rashidul Islam、Kamrun N. Keya、Shimei Pan 和 S. Pan。2020。《交叉性公平定义》。载于 *2020*年*IEEE*第*36*届国际数据工程会议（*ICDE*），1918–1921。数字对象标识符：https://doi.org/10.1109/ICDE48307.2020.00203。[18] Alexandra R. Givens 和 Meredith R. Morris。2020。《在算法公平、问责与透明度中聚焦残障视角》。载于 *FAT** *'20*。《style id='3'>2020年公平、问责与透明度会议论文集：2020年1月27日至30日，西班牙巴塞罗那。计算机协会，纽约，纽约州，684。数字对象标识符：https://doi.org/10.1145/3351095.3375686。[19] Leif Hancox-Li 和 I. E. Kumar。2021年。特征重要性方法中的认识论价值：女性主义认识论的启示。载于*2021*年*ACM*公平性、问责制与透明度会议论文集。ACM数字图书馆。计算机协会，美国纽约州纽约市，817–826。数字对象标识符：https://doi.org/10.1145/3442188.3445943。[20] Alex Hanna、Emily Denton、Andrew Smart 和 Jamila S mith-Loud。2020年。迈向算法公平中的批判性种族方法论。载于*FAT** *'20*。*2020*年公平、问责与透明度会议论文集：*2020*年*1*月*27*日至*30*日，西班牙巴塞罗那。计算机协会，纽约，纽约州，501–512。数字对象标识符：https://doi.org/10.1145/3351095.3372826。[21] Donna Haraway。1988年。情境知识：女性主义中的科学问题与局部视角的特权。《女性主义研究》 14, 3, 575–599。数字对象标识符：https://doi.org/10.2307/3178066。[22] Sandra Harding，编。1991年。谁的科学？谁的知识？ 从女性的生活出发思考。康奈尔大学出版社，伊萨卡，纽约州。[23] Sandra Harding。1992年。重新思考立场认识论：什么是“强客观性”？《百年评论》 36, 3, 437–470。[24] Sandra Harding。2008年。自下而来的科学。女性主义、后殖民性与现代性。下一波：女性研究的新方向。杜克大学出版社，达勒姆。

[25] Sandra Harding。2015年。客观性与多样性。科学研究的另一种逻辑。芝加哥大学出版社，芝加哥、伦敦。[26] Patricia Hill Collins。1990年。黑人女性主义思想：知识、意识与赋权政治。Routledge，纽约。[27] Patricia Hill Collins。2016年。作为对抗性知识的黑人女性主义思想。《批判性定性研究新方向》5, 3, 133–144。数字对象标识符：https://doi.org/10.1525/dcqr.2016.5.3.133。[28] Linus T.-L. Huang、Hsiang-Yun Chen、Ying-Tung Lin、Tsung-Ren Huang 和 Tzu-Wei Hung。2022年。改善算法偏见，或为何可解释人工智能需要女性主义哲学。《女性主义哲学季刊》8, 3。[29] Lilly Irani、Janet Vertesi、Paul Dourish、Kavita Philip 和 Rebecca E. Grinter。2010 年。后殖民计算。载于 *CHI 2010*。会议论文集：第*28*届年度*CHI*计算系统中的人为因素会议：论文与札记。ACM，纽约，纽约州，1311–1320。数字对象标识符：https://doi.org/10.1145/1753326.1753522。[30] Michael Katell、Meg Young、Dharma Dailey、Bernease Herman、Vivian Guetler、Aaron Tam、Corinne Bintz、Daniella Raz 和 P. M. Krafft。2020年。迈向算法公平的情境化干预。载于*FAT* '20*。*2020*年公平、问责与透明度会议论文集：*2020*年*1*月*27*日至*30*日，西班牙巴塞罗那。计算机协会，纽约，纽约州，45–55。数字对象标识符：https://doi.org/10.1145/3351095.3372874。[31] Goda Klumbytė、Claude Draude 和 Alex S. Taylor。2022年。机器学习的批判性工具：在机器学习系统设计中运用交叉性批判概念。载于*FAccT 2022*。*2022*年第*5*届*ACM*公平、问责与透明度会议：*2022*年*6*月*21*日至*24*日，韩国首尔。ICPS。计算机协会

计算机协会，纽约，纽约，1528–1541。数字对象标识符：https://doi.org/10.1145/3531146.3533207.[32] Hana Kopecká。2021。可解释人工智能：以人为中心的视角。WomENcourage，2021年9月22日至24日，布拉格。[33] Hana Kopecká 和 Jose Such。2020。面向文化心智的可解释人工智能。载于《面向人机交互的对话、解释与论证》研讨会。[34] Neha Kumar 和 Naveena Karusala。2019年。交叉性计算。交互 26，2，50–54。数字对象标识符：https://doi.org/10.1145/3305360。[35] Susan Leavy、Eugenia Siapera 和 Barry O'Sullivan。2021。人工智能伦理数据策展：一种基于女性主义认识论与种族批判理论的方法。载于*AIES '21*。*2021*年*AAAI/ACM*人工智能、伦理与社会会议论文集：*2021*年*5*月*19*日–*21*日，线上活动，美国。计算机协会，纽约，纽约州，695–703。数字对象标识符：https://doi.org/10.1145/3461702.3462598。[36] Sabelo Mhlambi。2020。从理性到关系性：乌班图作为人工智能治理的伦理与人权框架。卡尔中心讨论论文系列，2020-009。[37] Shakir Mohamed、Marie-Therese Png 和 William Isaac。2020。去殖民人工智能：去殖民理论作为人工智能领域的社会技术前瞻。《哲学与技术》 33，4，659–684。数字对象标识符：https://doi.org/10.1007/s13347-020-00405-8。[38] Sina Mohseni、Niloofar Zarei 和 Eric D. Ragan。2021。可解释人工智能系统设计与评估的多学科综述及框架。《*ACM*交互式智能系统汇刊》 11，3-4，1–45。数字对象标识符：https://doi.org/10.1145/3387166。[39] Claudia Müller-Birn、Katrin Glinka、Peter Sörries、Michael Tebbe 和 Susanne Michl。2021。情境化

以人为中心的解释用户界面设计案例研究。CHI 2021以人为中心的可解释人工智能研讨会（HCXAI）。[40] Milda Norkute。2021。人工智能可解释性：为何一种解释无法适用于所有人。CHI 2021以人为中心的可解释人工智能研讨会（HCXAI）。[41] Ihudiya F. Ogbonnaya-Ogburu、Angela D. Smith、Alexandra To 和 Kentaro Toyama。2020。面向人机交互的批判种族理论。载于*2020*年*CHI*计算系统中的人为因素会议论文集。ACM数字图书馆。计算机协会，纽约，纽约州，美国，1–16。数字对象标识符：https://doi.org/10.1145/3313831.3376392。[42] Kavita Philip、Lilly Irani 和 Paul Dourish。2012。后殖民计算。《科学、技术与人类价值》 37，1，3–29。数字对象标识符：https://doi.org/10.1177/0162243910389594。[43] Mari E. Ramler。2021。酷儿可用性。《技术传播季刊》 30，4，345–358。数字对象标识符：https://doi.org/10.1080/10572252.2020.1831614。[44] Ari Schlesinger、W. K. Edwards 和 Rebecca E. Grinter。2017。交叉性人机交互：通过性别、种族与阶级理解身份。载于*2017*年*CHI*计算系统中的人为因素会议论文集。ACM，纽约，纽约州，5412–5427。数字对象标识符：https://doi.org/10.1145/3025453.3025766。[45] Ari Schlesinger、Kenton P. O'Hara 和 Alex S. Taylor。2018。让我们谈谈种族。载于*2018*年*CHI*计算系统中的人为因素会议论文集。ACM会议。ACM，纽约，纽约州，1–14。数字对象标识符：https://doi.org/10.1145/3173574.3173889。[46] Deepa Singh、Michal Slupczynski、Ajit G. Pillai 和 Vinoth P. S. Pandian。2022。在全球南方语境中探讨可解释性在可解释人工智能中。CHI 2022以人为中心的可解释人工智能研讨会（HCXAI）。[47] Ronal Singh、Upol Ehsan、Marc Cheong、Mark O. Riedl 和 Tim Miller。2021年。LEx：机器学习解释层级的操作化框架。[48] Mona Sloane、Emanuel Moss、Olaitan Awomolo 和 Laura Forlano。参与并非机器学习设计问题的解决方案。载于 ICML'20：第37届国际机器学习会议论文集。[49] Harini Suresh、Rajiv Movva、Amelia L. Dogan、Rahul Bhargava、Isadora Cruxen、Angeles M. Cuba、Guilia Taurino、Wonyoung So 和 Catherine D'Ignazio。2022年。迈向交叉性女性主义与参与式机器学习：支持女性杀害反向数据收集的案例研究。载于 *FAccT 2022*。*2022*年第*5*届*ACM*公平性、问责制与透明度会议：*2022*年*6*月*21*日至*24*日，韩国首尔。ICPS。计算机协会，纽约州纽约市，667–678。数字对象标识符：https://doi.org/10.1145/3531146.3533132。[50] Ellen M. Voorhees。2021年。系统解释：一个警示故事。CHI 2021以人为中心的可解释人工智能研讨会（HCXAI）。[51] Zijie J. Wang、Alex Kale、Harsha Nori、Peter Stella、Mark E. Nunnally、Duen H. Chau、Mihaela Vorvoreanu、Jennifer Wortman Vaughan 和 Rich Caruana。2022年。可解释性，然后呢？编辑机器学习模型以体现人类知识与价值观。载于 *KDD '22*。第*28*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集：*2022*年*8*月*14*日至*18*日，美国华盛顿特区。计算机协会，纽约州纽约市，4132–4142。数字对象标识符：https://doi.org/10.1145/3534678.3539074。

[52] Meredith Whittaker、Meryl Alper、Cynthia L. Bennett、Sara Hendren、Liz Kaziunas、Mara Mills、Meredith R. Morris、Joy Rankin、Emily Rogers、Marcel Salas 和 Sarah Myers West。2019年。残疾、偏见与人工智能。研讨会报告。AI Now Institute，纽约。