揭开幽灵的真相：女性主义交叉性可解释人工智能与制图法——用于呈现隐形劳动的方法

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

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

当代通过人工智能实现的自动化，离不开大量幕后人类劳动，而这些劳动往往既不为人所见，也未得到充分报酬。隐形劳动（包括标注和维护工作）是当代人工智能系统不可或缺的一部分，因此，让用户认识到其作用仍然十分重要。我们认为，可以通过可解释人工智能（XAI）设计来实现这一点，尤其是女性主义交叉性可解释人工智能设计。我们提出采用源自女性主义交叉性研究的制图法，以呈现人工智能的系统性视角，并纳入与隐形劳动相关的人工智能维度。

CCS 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 •计算方法 • 机器学习

附加关键词与短语：女性主义交叉性，可解释人工智能，隐形工作，可解释人工智能设计

美国计算机协会参考文献格式：

Goda Klumbytė、Hannah Piehl 和 Claude Draude。2023。《揭示幽灵：女性主义交叉性可解释人工智能与制图法：核算隐形劳动的方法》。工作坊：自动化幕后：幽灵般的照护劳动、维护与干预，计算机系统中的人为因素会议 CHI ’23，2023 年 4 月 23—28 日，德国汉堡，共 6 页。

1 引言：可解释人工智能（XAI）与隐形劳动

人工智能的兴起使得解释算法决策的需求日益凸显，这一点也体现在学术界对可解释人工智能（XAI）的研究兴趣不断增长上 [1, 2]。尽管学者们对可解释性的定义尚未达成一致，但他们都承认该领域的研究仍存在空白。什么构成一个（良好的）解释，仍有待达成共识 [1, 3]，而且可解释人工智能领域的术语使用也表明，不同概念之间缺乏明确的区分。例如，可解释性与可理解性相互交织，且经常被交替使用，尽管学者们大多认同，可理解性更侧重于人类理解模型的能力，并提供促成这一过程的信息手段；而可解释性则被视为一种以模型为中心的概念，旨在提供易于理解的解释 [1, 4]。

可解释人工智能旨在让算法模型的功能易于理解，从而为其输出表现提供依据[1, 5]。可解释人工智能涉及“向谁解释？”这一问题，这表明理想情况下，需要识别并考虑所有利益相关者 [5]。利益相关者群体包括开发者、理论家、伦理学家和用户 [4], , 他们的认知因素、经验、信息需求各不相同，目标也各有不同，例如可信性、因果性、可迁移性、信息性、置信度、公平性、可访问性、交互性或隐私意识 [1, 6]。因此，几乎不可能创建一个能够满足所有可解释人工智能受众需求的模型。

隐形劳动是一个总括性术语，可能涉及后台工作（行政任务）、常规工作（需要问题解决技能和深厚知识）、由（社会意义上）不可见的人从事的工作（家务劳动），或非正式劳动（沟通、社交、情感劳动）[7]。女性主义学者运用隐形工作的概念，指出劳动（从字面和比喻意义上）如何被看见以及如何被评价，其中存在（往往相互交织的）性别、阶级和种族化不平等与分工[8, 9]。多位学者发现，在梳理工作场所的活动、任务和可供性时，只有显而易见的劳动才会被记录下来[8, 10, 11]。幽灵劳动，包括标注和维护工作，是当代人工智能系统不可或缺的一部分，因此，让用户认识到它所发挥的作用，并着重揭示隐形、价值被低估且薪酬偏低的劳动形式，至关重要。

在学术和专业语境中，可解释性被用作一个技术术语，意味着提供可解释的算法系统完全可以由技术专家单独处理。然而，当系统及其影响未被纳入考量时，仅仅解释模型的运作原理并不足够。相反，为了设计可解释人工智能系统并防止算法偏见再现，需要纳入多元的利益相关者群体和其他学科的知识[12]。我们认为，这种对可解释性的广义理解，是将可解释人工智能视为一个领域的概念基础；这一领域能够帮助人机交互和人工智能形成方法与路径，展现人工智能并非理想化的技术奇迹[13] ，而是一个复杂的技术组装体和基础设施，其中包含人类劳动和复杂的权力动态。然而，我们认为，要充分实现这一点，可解释人工智能还需要借鉴女性主义认识论立场，考量可解释人工智能过程中知识生产的背景与情境性。

2 女性主义认识论立场 **NS**

交叉性揭示了一个人被归属或自我认同的社会类别如何相互交织，并加剧其遭受歧视或享有特权的经历。除了种族、阶级和性别之外，许多其他社会类别——例如宗教、性取向、地域、残障/能力——也会影响社会、文化和经济资源以及权力观念。借助十字路口这一隐喻，交叉性作为一种分析工具，旨在凸显多样且因情境而异的歧视经历，并展现社会类别如何相互作用、彼此影响[14]。

女性主义认识论认为，知识总是处于特定情境之中，这意味着不存在所谓普遍、客观或中立的知识 [15]。与社会类别类似，（科学）知识也与社会和文化背景交织在一起，从而形成具有局限性、主观性，因而处于特定情境之中的知识实践。因此，情境化知识这一概念主张，应通过“将局部视角与暂歇的声音汇聚为一个集体主体位置，从而展望有限的具身性得以延续的方式，在限度与矛盾之中生活——从某个地方出发的视角”[15, 第590]页，以更具多元性的方式理解知识。立场理论阐释了一个人的立场如何影响（科学）知识生产与理解实践[16, 17, 18, 19, 20, 21]。将知识视为处于特定情境之中，意味着考量权力关系，以及结构性、认识论和系统性暴力的记录。通过将边缘化视角置于中心，并关注鲜为人知的历史与另类知识实践，往往不可见的压迫形式便能显现出来，并得到关注。立场理论不仅呼吁承认

知识所处的具体社会、文化与历史位置，也主张利用这种局部性来抵制偏见与歧视的再生产。

3 女性主义交叉性可解释人工智能

女性主义认识论视角通过质疑并重新审视若干方面，拓展了可解释人工智能的框架。首先，它要求始终在特定的历史、政治和社会文化语境中理解并设计可解释性。此外，由于女性主义视角关注相互交织的结构性不平等，这一语境不仅限于直接的社会技术应用场景，还包括相关人工智能系统赖以运行的更广泛框架。系统边界的具体划定取决于个案，但女性主义视角要求以更具结构性、系统性且立足具体情境的方式理解系统 [22]。与较狭义的可解释人工智能技术理解不同，这种视角有助于采取整合的可解释人工智能方法 [12] ，并为将无形劳动——数据标注、系统维护、基础设施支持——纳入需要解释的系统运作范围提供依据。

其次，女性主义交叉性要求密切关注权力动态，并在知识生产和设计实践中以边缘化群体的视角为中心。从这一角度来看，权力动态具体涉及提出以下问题：人工智能系统让谁受益？谁又因此遭到剥削或被剥夺？谁在向谁解释这个系统，又是出于什么目的？此外，整个设计过程中都应持续提出这些权力问题；在此过程中，我们有一项规范性义务，即努力实现更公平、更正义的结果，并优先考虑处于较弱势地位者的视角。由于机械土耳其人等被隐形化的劳动形式往往报酬低廉，且受地缘政治不平等的影响，[23], 女性主义可解释人工智能为这类劳动以及劳动者的视角提供了可能，使其不仅得到关注，还能被优先纳入解释生成的立场之中。

第三，女性主义认识论立场主张将问责问题融入知识生产与设计实践。此处的问责并非局限于狭义地追问谁负有法律责任、出问题时谁来承担责备，而是积极努力，旨在提升回应能力：回应之能[15, 24, 25,26]。这种回应之能要求以一种方式构建可解释人工智能解决方案，促使利益相关者能够批判性地参与并回应相关人工智能系统。从女性主义视角看，系统的定义本就比某个特定的机器学习模型更为宽泛，因此，被隐形化的劳动者也可以被视为重要的利益相关者群体。

总而言之，女性主义交叉性可解释人工智能通过引导该领域关注情境化、重视权力关系并将受压制的视角置于中心，可以开辟一种路径，帮助我们核算《幽灵工作》和被隐形化的劳动形式，并生成不只解释相关人工智能所作具体决策，还涵盖系统运行方式及其与所处情境之纠缠的解释。我们认为，在方法论上，进一步借鉴女性主义交叉性研究，寻找处理这一更广泛生态系统层面的范例，也会有所助益。

4 制图法作为方法

我们提出，制图法是一种有用的方法，可以呈现人工智能的系统性视角，并纳入与不可见劳动相关的人工智能维度。具体而言，这里所说的制图，是女性主义文化研究、女性主义哲学[27, 28, 29, 30, 31] 和科学技术研究[32, 33], 中使用的制图，意指借助一组导航概念（例如性别），对特定现象进行绘图或追踪。其核心在于位置性，即从特定的学科、政治和文化立场出发进行创作。它突显具身的交叉性视角，这种视角不仅关注人的身体，也关注技术及其赖以维系的基础设施等物质实体，并将受支配的

视角置于中心[27]。因此，制图可以成为交叉性工具，以纳入不可见的劳动形式（数据收集、标注、维护等）作为人工智能系统不可或缺的组成部分，从而提高人们（包括可解释人工智能设计师和最终用户）对这类劳动、照护劳动及其权力动态的认识，并了解它们如何构成人工智能系统运行的基础。因此，多面向的制图可被视为女性主义可解释人工智能的合适方法。

制图可以是文本形式、视觉形式，或兼具两者。一个关注并凸显隐形劳动的人工智能系统制图案例，是 Crawford 和 Joeler 的项目“人工智能解剖” [34]，，该项目概述了驱动智能音箱所需的人力、自然和技术资源、知识与运作。在我们先前的工作中 [35]，我们运用机器学习系统的衍射映射，捕捉并理解该系统可能产生的影响，其中包括对权力关系的分析。在我们的映射练习中，一组人机交互从业者着重通过构建、扰动和干预之间的关系来理解这些影响（图1展示了一个例子）。我们请参与者仔细考察系统的基础设施层面及其与社会环境的交互，以指出他们所分析系统中的这些关系。我们特别鼓励他们探究社会、技术或学科相关要素、话语或价值要素，以及系统的运行逻辑。

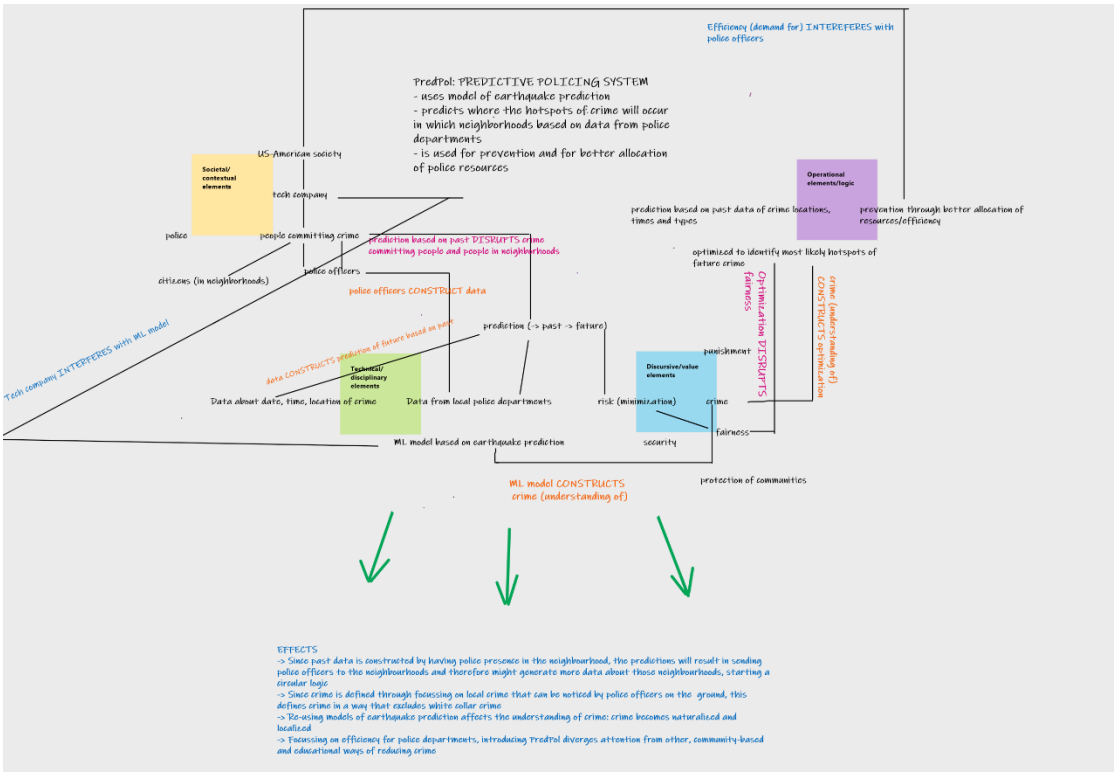

图1：预测性警务系统 PredPol 的衍射映射。

虽然制图法绝非应对人工智能基础设施及其隐形劳动形式的唯一方法工具，但我们认为，它可以成为一个良好的起点，尤其能帮助可解释人工智能设计师开始思考系统与社会技术情境、权力和劳动之间的交互问题。

5 讨论

在这篇立场论文中，我们提出，女性主义交叉性可解释人工智能与制图法能够帮助我们认识到支撑人工智能系统运行的被隐形化的劳动形式。希望这也能让我们更充分地理解人工智能系统，并促成更有成效的批判性参与。我们还需要进一步研究如何在实践中将这一概念视角付诸实施：它在多大程度上可以付诸实施？它是否更适合作为一套用于问题定义和提高设计师敏感度的指导原则？谁能够将其付诸实施？实施这样一种更具系统性、更广泛的可解释人工智能视角，需要哪些资源和技能？为此，我们认为，尝试采用社会科学和人文学科的跨学科方法论十分重要；同时还应探讨如何培养人机交互研究者以及人工智能系统用户的集体回应能力，使他们能够更积极地参与批判性思考。

致谢

本研究得到大众汽车基金会“人工智能与未来社会”项目的资助，是协作项目“人工智能取证：通过可解释性实现视觉人工智能系统问责”的一部分。

参考文献

[1] Alejandro Barredo-Arrieta、Natalia Díaz-Rodríguez、Javier Del Ser、Adrien Bennetot、Siham Tabik、Alberto Barbado、Salvador Garcia、Sergio Gil- Lopez、Daniel Molina、Richard Benjamins、Raja Chatila 和 Francisco Herrera。2020。可解释人工智能（XAI）：概念、分类体系、机遇与挑战，迈向负责任的人工智能。Information Fusion 58，82–115。DOI: https://doi.org/10.1016/j.inffus.2019.12.012。[2] Prashan Madumal、Ronal Singh、Joshua Newn 和 Frank Vetere。2018。面向可解释人工智能的交互设计。载于《第30届澳大利亚计算机与人类交互会议论文集》。美国计算机协会，纽约州纽约市，607–608。DOI: https://doi.org/10.1145/3292147.3293450。[3] Roberto Confalonieri、Ludovik Coba、Benedikt Wagner 和 Tarek R. Besold。2021。可解释人工智能的历史视角。Wiley interdisciplinary reviews. Data mining and knowledge discovery 11，1。DOI: https://doi.org/10.1002/widm.1391。[4] Alun Preece、Dan Harborne、Dave Braines、Richard Tomsett 和 Supriyo Chakraborty。2018。可解释人工智能中的利益相关者。[5] 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》。施普林格，尚，449–466。DOI: https://doi.org/10.1007/978-3-030-60117-1_33。[6] Harmanpreet Kaur、Eytan Adar、Eric Gilbert 和 Cliff Lampe。2022。明智人工智能：运用意义建构理论重新构想可解释性与可说明性。载于 FAccT 2022。《2022年第五届美国计算机协会公平、问责与透明度会议：2022年6月21日至24日，韩国首尔》。ICPS。美国计算机协会，纽约州纽约市，702–714。DOI: https://doi.org/10.1145/3531146.3533135。[7] Bonnie A. Nardi 和 Yrjö Engeström。1999。风中的网络：隐形工作的结构。计算机支持的协同工作 8，1-2，1–8。DOI：https://doi.org/10.1023/A:100869462128。[8] Marion G. Crain、Winifred Poster 和 Miriam A. Cherry 编。2016。隐形劳动：当代世界中的隐秘工作。加利福尼亚大学出版社，奥克兰，加利福尼亚州。[9] Lucy Suchman。1995。让工作变得可见。Communications of the ACM 38，9，56–64。DOI：https://doi.org/10.1145/223248.223263。[10] Christina Courtright。2007。信息行为研究中的情境。Annual Review of Information Science and Technology 41，1，273–306。DOI: https://doi.org/10.1002/aris.2007.1440410113。[11] Geoffrey C. Bowker 和 Susan L. Star。1999。梳理事物：分类及其后果。Inside technology。麻省理工学院出版社，剑桥。[12] Linus Ta-Lun Huang、Hsiang-Yun Chen、Ying-Tung Lin、Tsung-Ren Huang 和 Tzu-Wei Hung。2022。缓解算法偏见，或为何可解释人工智能需要女性主义哲学。Feminist Philosophy Quarterly 8，3。[13] M. C. Elish 和 Danah Boyd。2018。在大数据与人工智能的魔力中定位方法。Communication Monographs 85，1，57–80。DOI: https://doi.org/10.1080/03637751.2017.1375130。[14] Kimberle Crenshaw。1989。消除种族与性别交叉问题的边缘化：对反歧视学说、女性主义理论和反种族主义政治的黑人女性主义批判。University of Chicago Legal Forum 1989，8。[15] Donna Haraway。1988。情境化知识：女性主义中的科学问题与局部视角的特权。Feminist Studies 14，3，575–599。DOI：https://doi.org/10.2307/3178066。[16] Patricia Hill Collins。2016。作为对抗性知识的黑人女性主义思想。Departures in Critical Qualitative Research 5，3，133–144。DOI: https://doi.org/10.1525/dcqr.2016.5.3.133。[17] Sandra Harding。2015。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。[18] Sandra Harding。2008。自下而来的科学：女性主义、后殖民性与现代性。Next wave: New directions in women's studies。杜克大学出版社，达勒姆。

[19] Sandra Harding。1992。重新思考立场认识论：什么是“强客观性”？《世纪评论》36，3，437–470。[20] Sandra Harding。1991。谁的科学？谁的知识？载于《谁的科学？谁的知识？从女性的生活出发思考》，Sandra Harding 编。康奈尔大学出版社，伊萨卡，纽约州。[21] Patricia Hill Collins。1990。黑人女性主义思想：知识、意识与赋权政治。劳特利奇出版社，纽约。[22] Claude Draude、Goda Klumbytė、Phillip Lücking 和 Pat Treusch。2020。情境化算法：一种针对偏见的社会技术系统方法。《在线信息评论》44，2，325–342。DOI：https://doi.org/10.1108/OIR-10-2018-0332。[23] Mary Gray 和 Siddharth Suri。2019。《幽灵工作：亚马逊、谷歌和优步如何创造新的全球底层阶级》。霍顿·米夫林·哈考特出版公司，波士顿。[24] Claude Draude。2020。“边界并非静止不动”：在人机交互中从交互转向能动性内交互。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，包括互联网／网络与人机交互，12181。施普林格，20–32。DOI：https://doi.org/10.1007/978-3-030-49059-1_2。[25] Karen Barad。2007。《与宇宙相遇于中途：量子物理学与物质和意义的纠缠》。杜克大学出版社，达勒姆。[26] Karen Barad。2012。“内交互”（Adam Kleinmann 对 Karen Barad 的访谈）。《Mousse》，76–81。[27] Rosi Braidotti。2021。后人类女性主义与性别方法论。载于《为何谈性别？》，Jude Browne 编。剑桥大学出版社，101–125。DOI：https://doi.org/10.1017/9781108980548.007。[28] Rick Dolphijn 和 Iris van der Tuin。2012。新唯物主义：访谈与制图。开放人文出版社。[29] Rosi Braidotti。2011a。游牧主体：当代女性主义理论中的具身性与性差异。哥伦比亚大学出版社，纽约。[30] Rosi Braidotti。2011b。游牧理论：便携版 Rosi Braidotti。性别与文化。哥伦比亚大学出版社，纽约。[31] M. Jacqui Alexander 和 Chandra T. Mohanty。2010。知识与权力的制图：跨国女性主义作为激进实践。载于《批判性跨国女性主义实践》，Amanda Lock Swarr 和 Richa Nagar 编。纽约州立大学系列丛书：实践——行动中的理论。纽约州立大学出版社，奥尔巴尼，23–45。[32] Lea Skewes 和 Stine W. Adrian。2018。认识论、行动主义与纠缠——重新思考知识生产。《女性、性别与研究》，1，15–31。DOI：https://doi.org/10.7146/kkf.v27i1.109677。[33] Donna Haraway。2016。《与麻烦共处：在克苏鲁世建立亲缘关系》。实验性未来。杜克大学出版社，达勒姆、伦敦。[34] Kate Crawford 和 Vladan Joler。2018。《人工智能系统解剖图》（2018）。检索于2023年2月21日，网址：https://anatomyof.ai/。[35] Goda Klumbytė、Claude Draude 和 Alex S. Taylor。2022。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年第五届美国计算机协会公平、问责与透明度会议：2022年6月21日至24日，韩国首尔》。ICPS。美国计算机协会，纽约州纽约市，1528–1541。DOI：https://doi.org/10.1145/3531146.3533207。