arXiv:2305.00739v1 [cs.HC] 1 May 2023

生成以过程为中心的解释，以促进算法决策中的可争议性：挑战与机遇

Mireia Yurrita
代尔夫特理工大学，代尔夫特，荷兰
m.yurritasemperena@tudelft.nl

Agathe Balayn
代尔夫特理工大学，代尔夫特，荷兰
a.m.a.balayn@tudelft.nl

Ujwal Gadiraju
代尔夫特理工大学，荷兰代尔夫特
u.k.gadiraju@tudelft.nl

特此授权将本作品的部分或全部内容制作成数字版或纸质副本，供个人或课堂使用，无需付费，但前提是不得为营利或商业目的制作或分发副本，且副本首页须附上本声明及完整引文。本作品中第三方组件的版权必须得到尊重。其他用途请联系所有者/作者。版权归所有者/作者所有。*CHI'23,*, 2023年4月23日至28日，德国汉堡 ACM
978-1-4503-6819-3/20/04.https://doi.org/10.1145/3334480.XXXXXXX

摘要

人类—人工智能决策正变得日益普遍，而解释已被提出用于促进更好的人类—人工智能交互。近期研究探讨了解释对算法决策中决策主体公平性感知的积极影响。尽管取得了这些进展，大多数研究仍孤立地考察解释的作用，将解释视为目的本身，并将其简化为通过可解释人工智能方法提供的技术解决方案。在这篇愿景论文中，我们认为，解释对公平性感知的影响应当结合决策主体对这类决策提出争议 的权利来考察。由于可争议人工智能系统在其整个生命周期中都允许人工干预，可争议性 要求解释不应局限于结果，还应阐明最初促成算法系统开发与部署的理由。我们将此类解释称为以过程为中心的解释。本文引入以过程为中心的解释这一概念，并阐述生成和评估此类解释面临的一些主要挑战与研究机遇。

引言

算法系统日益被用于高风险决策流程中（例如，授予人们获取教育、就业和公共服务福利的流程中

正义
研究其各个维度上的公平性感知的多维构念 [8]。

程序正义
旨在捕捉人们对决策过程的公平性感知的正义维度（即程序公平感知） [8]。

事后可争议性补救与申诉程序 [26]。

设计中的可争议性围绕算法系统开发和部署过程中所做选择的讨论 [2]

[13]）。据称，这些系统有望带来广泛的社会经济效益 [14]。然而，人们对算法系统可能导致歧视的方式提出了诸多担忧 [33,6]。因此，近期研究开发了具备公平性意识的算法系统，以确保结果分配的公平性（即分配正义） [11, 17]。然而，即使算法系统符合某种规范性标准而具有公平性，人们也未必会认为其公平。关于人类决策的法律和组织心理学文献早已表明，决策主体不仅关心结果的分配，也关心促成这些结果的流程（即程序正义） [45, 25]。鉴于解释在揭示算法决策如何作出的过程中所发挥的作用，多项人机交互（HCI）研究 [10, 5] 探讨了解释对决策主体公平性感知的影响。

尽管现有研究作出了重要贡献，但仍未能克服两个主要局限。首先，如果解释没有同时提供对决策提出异议的选项，就可能无法有效地让决策主体发声——而这是 Thibaut 和 Walker 所定义的程序正义的一个关键方面 [45]。正如法律学者在讨论解释权是否存在时所指出的，在欧盟《通用数据保护条例》的背景下（例如， [40, 43]），应结合决策主体对自动化决策提出异议的能力来研究解释。可争议性已被视为程序正义的一种形式，能够提升人们对公平性的感知 [2]。反过来，可争议性也需要解释，以便决策主体提出有力的论据 [40]。

其次，解释一直局限于由技术性可解释人工智能（XAI）方法实现的结果解释。然而，主观酌情选择

在算法系统的整个开发和部署过程中所作出的选择会影响最终的算法行为 [7]„ 而决策主体可能希望对这些行为提出质疑 [2]。因此，有必要将解释扩展至系统的整个生命周期，并从事后可争议性 转向 设计中的可争议性 [2] 。

在这篇愿景论文中，我们认为，应结合决策主体质疑自动化决策的能力，研究解释对其公平性感知的影响。我们提出并阐释以过程为中心的解释，认为这是实现设计中的可争议性所必需的要素 [2]。以过程为中心的解释记录算法系统在开发和部署过程中作出的决策（例如有关训练数据的信息）。然而，与以数据为中心的解释 [3],等概念不同，以过程为中心的解释还会提炼这些选择背后的理由，以便接受公开审查。在接下来的章节中，我们将说明为何需要以过程为中心的解释、此类解释的性质，以及生成这些解释所面临的挑战和研究机遇。我们扩展了 Schoeffer 等人 [42] 在 ACM CHI 2022以人为中心的可解释人工智能研讨会（HCXAI）上提出的概念框架，将以过程为中心的解释纳入程序公平、依赖与分配公平之间互动关系的关键要素（见附录图 1）。这一框架与 Alfrink 等人 [2]提出的设计中的可争议性框架相辅相成。

以过程为中心的解释

需要以过程为中心的解释。
在实际应用中，算法系统往往被视为用于决策的“客观”手段。然而，不同利益相关者在整个过程中会做出若干“未公开却影响深远”的 [7] 主观自由裁量选择，贯穿于——

反思性
一种审视机器学习从业者（例如数据工程师和数据科学家）视为理所当然的假设和日常流程的实践 [29, 20, 7, 12]。

协作性反思实践

利益相关者与学科之间的冲突得以转化为反思的过程 [20]。

机器学习（ML）系统的开发和部署过程 [35, 36, 39]。例如，要将高度复杂的战略业务目标转化为代表可处理工程问题的设计要求，就需要确定恰当的目标变量，或这些变量的代理变量 [34]。在数据方面，各种价值取向和决策会影响数据的收集方式（例如，关于纳入哪些数据、排除哪些数据，以及如何衡量数据质量的决策） [37], 、转换方式（例如，以异常值的形式移除少数类数据所带来的影响） [4], 、分析方式 [31] 和可视化方式 [23]。这些活动中根深蒂固的主观性 [1, 44] 常被视为导致下游有害影响的根源 [39]。以过程为中心的解释依托反思性，将规范性假设转化为有意义的信息 [40]，从而促进设计中的可争议性 [2]。

以过程为中心的解释的特征。以过程为中心的解释借鉴并拓展了诸如以数据为中心的解释 [3],*Datasheets* [15], 数据集营养标签 [21] 或模型卡 [30]等概念。以数据为中心的解释包含有关训练数据的信息：数据收集机制（即收集的数据量、数据来源、数据收集者和标注流程）、人口统计信息（即数据样本中性别、种族、年龄和国家分布的信息）、建议用途以及潜在问题。与以数据为中心的解释类似，以过程为中心的解释超越了结果透明度，揭示了促成相关算法系统开发的过程细节。然而，我们认为，以过程为中心的解释是赋能决策对象行使其对自动化决策提出异议的权利的一种途径。我们呼应 Selbst 等人 [43] 的观点，倡导能够使决策对象采取行动的功能性解释。就实践而言，与以数据为中心的解释不同——

解释，这意味着以过程为中心的解释着重揭示机器学习系统生命周期中各项自由裁量选择背后的理由。例如，与其仅仅描述数据的收集方式（以数据为中心的解释），以过程为中心的解释还会阐述选择某一数据来源的原因，并反思这一特定选择可能造成的潜在危害（例如，反思不将某些子群体纳入数据的理由 [32]）。明确这些理由，能够以正当理由（“为什么”）补充技术描述（“如何”） [40, 19]，并为围绕数据和建模选择展开辩论奠定基础 [2]。与 Datasheets [15]、数据集营养标签 [21]和模型卡 [30],不同，以过程为中心的解释面向决策主体（或其法律代表 [26]），而非数据集或模型的创建者和使用者。也就是说，以过程为中心的解释应向决策主体提供“更高层次或更具体的信息” [40]，使其能够在申诉过程中进行理性且富有成效的辩论式交流。例如，这类解释包括反映组织价值观的决策（例如，促使组织优先选择不透明系统而非准确性的基本选择）。

生成和评估以过程为中心的解释：挑战与机遇。

要在整个机器学习流程中有效促进反思性实践并生成以过程为中心的解释，面临若干挑战，也伴随着相应的研究机遇。我们从三个主要抽象层次对这些挑战与机遇进行了识别和综合。

微观层面的挑战（**个体**）。在机器学习从业者中培养反思性实践，主要挑战之一是提高他们对自由裁量选择的主观性质的认识

，以及这些选择可能造成的下游影响。例如，在数据整理、数据策展和特征工程过程中，数据科学家可能需要决定如何处理数据，而这反过来可能引入偏见并导致歧视 [32]。这一抽象层面上的挑战包括：（1）开发适当的基础设施，以及（2）为机器学习从业者提供进一步的培训和指导，使他们能够采用反思性实践。就技术基础设施而言，先前的研究指出，有必要开发用于数据版本控制 [41]的技术接口，或用于追踪数据转换如何影响公平性指标 [4]的技术接口。至于如何培训从业者开展反思性实践，关于公平性工具包等反思工具的研究指出，这些工具缺乏可操作的指导 [9]，并且需要进一步支持从业者反思其自由裁量选择可能造成的下游危害 [28, 27]。这包括提供资源，让从业者通过真实案例 [9, 22, 24]和入职培训材料 [24],练习反思，或在反思工具中以向导的形式整合更有针对性的指导 [4]。需要指出的是，即便提供了 Datasheets [15] 形式的指导，许多人仍难以明确理解这些反思性文档与负责任人工智能实践之间的联系 [18]。因此，有必要明确阐释这些联系。

中观层面的**协商**。算法系统如何运作并非既定不变，而是在问题制定阶段 [36]协商确定的。同样，如何界定和操作化一个问题，也总是经过协商的 [32]。鉴于协商过程贯穿整个机器学习流程，已有研究指出，有必要促进协作性反思实践。协作性反思实践反过来需要借助工具来促进沟通 [9, 22, 38]，并需要数字对象（例如，为非工程师量身定制的数据可视化和报告），以推动话语协商。

化和报告 [9, 24]），以推动话语协商 [20]。

宏观层面的挑战（**治理**）。组织层面的倡议和政策对于有效促进从业者开展反思性实践并推动生成以过程为中心的解释至关重要。以往研究指出，在要求个人和团队参与反思性实践时，必须考虑业务需求 [27]和组织约束 [16]。因此，实施全组织范围的策略（例如，提供资源 [27] 和支持 [28]）的重要性不容忽视。这些策略包括但不限于：在利益相关者之间建立对质量目标的共同理解、改进跨组织文档记录，以及建立激励机制，鼓励利益相关者参与反思性实践 [44]。

可以通过记录从这些反思性实践中提炼出的理由来生成以过程为中心的解释。从记录到以过程为中心的解释，还会带来一些额外的挑战。首先，需要界定哪些内容可以提出争议 [26]，并据此针对与可争议性相关的信息定制以过程为中心的解释。其次，应以满足决策主体（或其法律代表 [26]）需求的方式呈现这些信息。第三，还应制定评估协议，以确保以过程为中心的解释能够有效实现可争议性。我们需要探索如何跨不同层面有效且全面地评估以过程为中心的解释。例如，这些评估协议可能要求迭代更新以过程为中心的解释。如何统筹并落实此类协议，是一个有前景的未来研究方向。

致谢

本研究部分得到欧盟地平线2020研究与创新计划的资助，资助依据为玛丽·斯克沃多夫斯卡-居里资助协议第955990号，并得到代尔夫特理工大学人工智能倡议的支持。

参考文献

[1] Andrea Aler Tubella、Flavia Barsotti、Rüya Gökhan Koçer 和 Julian Alfredo Mendez。2022。公平性干预的伦理影响：工程选择背后可能隐藏着什么？《伦理学与信息技术》24, 1（2022年3月），12。DOI：http://dx.doi.org/10.1007/s10676-022-09636-z

[2] Kars Alfrink、Ianus Keller、Gerd Kortuem 和 Neelke Doorn。2022。通过设计实现可争议的人工智能：迈向一个框架。《心灵与机器》（2022年8月）。DOI：http://dx.doi.org/10.1007/s11023-022-09611-z

[3] Ariful Islam Anik 和 Andrea Bunt。2021。以数据为中心的解释：解释机器学习系统的训练数据以促进透明度。载于*2021*年*CHI*计算机系统中的人为因素会议论文集。ACM，纽约，纽约州，美国。DOI：http://dx.doi.org/10.1145/3411764.3445736

[4] Agathe Balayn、Christoph Lofi 和 Geert-Jan Houben。2021 年。决策支持中数据偏差与不公平的管理：关于在数据管理和分析系统中识别并缓解偏差与不公平的机器学习和数据工程方法综述。《*VLDB*期刊》30，5（2021 年 9 月），739–768。DOI：http://dx.doi.org/10.1007/s00778-021-00671-8

[5] Reuben Binns、Max Van Kleek、Michael Veale、Ulrik Lyngs、Jun Zhao 和 Nigel Shadbolt。2018 年。'这是“将人类简化为百分比”；算法决策中的正义观感。（1 2018）。DOI：http://dx.doi.org/10.1145/3173574.3173951

[6] Joy Buolamwini 和 Timnit Gebru。2018 年。性别阴影：商业性别分类中的交叉性准确率差异。载于第 *1* 届公平性、问责制与透明度会议论文集（《机器学习研究论文集》），Sorelle A Friedler 和 Christo Wilson（编），第 81 卷。PMLR，77–91。https: // proceedings.mlr.press/v81/buolamwini18a.html

[7] Scott Allen Cambo 和 Darren Gergle。2022 年。模型立场性与计算反思性：促进数据科学中的反思性。在 *CHI* 人机交互系统中的人为因素会议。ACM，纽约州纽约，美国，1–19。DOI：http://dx.doi.org/10.1145/3491102.3501998

[8] Russell Cropanzano。2012 年。《工作场所中的公正：从理论到实践》。第 2 卷。

[9] Wesley Hanwen Deng、Manish Nagireddy、Michelle Seng Ah Lee、Jatinder Singh、ZhiweiSteven Wu、Kenneth Holstein 和 Haiyi Zhu。2022 年。探索机器学习从业者（尝试）如何使用公平性工具包。在*2022*年 *ACM* 公平性、问责制与透明度会议。ACM，纽约州纽约，美国，473–484。DOI：http://dx.doi.org/10.1145/3531146.3533113

[10] Jonathan Dodge、Q. Vera Liao、Yunfeng Zhang、Rachel K. E. Bellamy 和 Casey Dugan。2019年。《模型解释：解释如何影响公平性判断的实证研究》。（2019年1月）。DOI：http://dx.doi.org/10.1145/3301275.3302310

[11] Cynthia Dwork、Moritz Hardt、Toniann Pitassi、Omer Reingold 和 Richard Zemel。2012。通过认知实现公平。在第三届理论计算机科学创新会议（*ITCS '12*）论文集。美国计算机协会，美国纽约州纽约，214–226。DOI：http://dx.doi.org/10.1145/2090236.2090255

[12] M. C. Elish 和 Tarana Boyd。2018。将方法置于大数据与人工智能的魔力之中。《传播学专论》85, 1（2018 年第 1 期），57–80。DOI：http://dx.doi.org/10.1080/03637751.2017.1375130

[13] 欧盟委员会。2021。《关于制定人工智能协调规则（人工智能法案）并修订若干欧盟立法法案的条例提案》。[14] Luciano Floridi。2019。《将数字伦理原则转化为实践：不道德行为的五种风险》。《哲学与技术》32, 2（2019 年第 6 期）。DOI：http://dx.doi.org/10.1007/s13347-019-00354-x[15] Timnit Gebru、Jamie Morgenstern、Briana Vecchione、Jennifer Wortman Vaughan、Hanna Wallach、Hal Daumé III 和 Kate Crawford。2021。数据集数据说明书。《*ACM* 通讯》64, 12（2021 年第 12 期），86–92。DOI：http://dx.doi.org/10.1145/3458723[16] Ben Green。2021。《科技伦理的争议：实践中技术伦理的社会技术方法》。《社会计算期刊》2, 3（2021 年第 9 期），209–225。DOI：http://dx.doi.org/10.23919/JSC.2021.0018[17] Moritz Hardt、Eric Price 和 Nathan Srebro。2016。《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论文集（*NIPS'16*）. Curran Associates Inc.，美国纽约州雷德胡克，3323–3331。[18] Amy Heger、Elizabeth B. Marquis、Mihaela Vorvoreanu、Hanna Wallach 和 Jennifer Wortman Vaughan。2022。了解机器学习从业者对数据文档的认知、需求、挑战与期望。（2022年6月）。[19] Clément Henin 和 DanielLe Métayer。2021。超越可解释性：算法决策系统的可辩护性与可争议性。《人工智能与社会》（2021年7月）。DOI：http://dx.doi.org/10.1007/s00146-021-01251-8[20] Simon David Hirsbrunner、Michael Tebbe 和 Claudia Müller-Birn。2022。从批判性技术实践到反思性数据科学。《融合：新媒体技术研究国际期刊》（2022年11月），135485652211322。DOI：http://dx.doi.org/10.1177/13548565221132243[21] Sarah Holland、Ahmed Hosny、Sarah Newman、Joshua Joseph 和 Kasia Chmielinski。2018。数据集营养标签：推动提升数据质量标准的框架。（2018年5月）。[22] Kenneth Holstein、Jennifer Wortman Vaughan、Hal Daumé、Miro Dudik 和 Hanna Wallach。2019。提升机器学习系统的公平性。载于*2019*年*CHI*计算机系统中的人为因素会议论文集。ACM，纽约，美国，1–16。DOI：http://dx.doi.org/10.1145/3290605.3300830[23] S Kandel、A Paepcke、J M Hellerstein 和 J Heer。2012。企业数据分析与可视化：一项

访谈研究。*IEEE* 可视化与计算机图形学汇刊18, 12（2012年12月），2917–26。DOI：http://dx.doi.org/10.1109/TVCG.2012.219[24] Michelle Seng Ah Lee 和 Jat Singh。2021。开源公平性工具包的现状与不足。载于*2021*年*CHI*计算机系统中的人为因素会议论文集。ACM，纽约，美国，1–13。DOI：http://dx.doi.org/10.1145/3411764.3445261[25] Gerald S. Leventhal。1980。公平理论该如何处理？载于社会交换。Springer 美国，波士顿，马萨诸塞州，27–55。DOI：http://dx.doi.org/10.1007/978-1-4613-3087-5{_}2[26] Henrietta Lyons、Eduardo Velloso 和 Tim Miller。2021。概念化可争议性：对算法决策提出异议的视角。（2021年2月）。DOI：http://dx.doi.org/10.1145/3449180[27] Michael Madaio、Lisa Egede、Hariharan Subramonyam、Jennifer Wortman Vaughan 和 Hanna Wallach。2022。评估人工智能系统的公平性：人工智能从业者的流程、挑战与支持需求。*ACM* 人机交互论文集6，计算机支持的协同工作1（2022年3月），1–26。DOI：http://dx.doi.org/10.1145/3512899[28] Michael A. Madaio、Luke Stark、Jennifer Wortman Vaughan 和 Hanna Wallach。2020。共同设计检查清单，以了解人工智能公平性方面的组织挑战与机遇。载于*2020*年*CHI*计算机系统中的人为因素会议论文集。ACM，纽约，美国，1–14。DOI：http://dx.doi.org/10.1145/3313831.3376445[29] Milagros Miceli、Tianling Yang、Laurens Naudts、Martin Schuessler、Diana Serbanescu 和 Alex Hanna。2021. 记录计算机视觉数据集。载于*2021* 年 *ACM* 公平性、问责制与透明度会议论文集。ACM，纽约州纽约，美国，第 161–172 页。DOI：http://dx.doi.org/10.1145/3442188.3445880

[30] Margaret Mitchell、Simone Wu、Andrew Zaldivar、Parker Barnes、Lucy Vasserman、Ben Hutchinson、Elena Spitzer、Inioluwa Deborah Raji 和 Timnit Gebru。2018 年。模型报告模型卡。（2018 年 10 月）。DOI：http://dx.doi.org/10.1145/3287560.3287596

[31] Michael Muller、Ingrid Lange、Dakuo Wang、David Piorkowski、Jason Tsay、Q. Vera Liao、Casey Dugan 和 Thomas Erickson。2019年。《数据科学工作者如何与数据协作》。载于*2019*年 *CHI* 人机交互系统中的人为因素会议论文集。ACM，纽约州纽约，美国，1–15页。DOI：http://dx.doi.org/10.1145/3290605.3300356[32] Michael Muller 和 Angelika Strohmayer。2022。数据科学中的遗忘实践。载于*CHI* 人机交互系统中的人为因素会议。ACM，纽约州纽约，美国，1–19页。DOI：http://dx.doi.org/10.1145/3491102.3517644[33] Cathy O’neil。2016。数学毁灭性武器：大数据如何加剧不平等并威胁民主。Broadway Books。[34] Samir Passi 和 Solon Barocas。2019。问题构建与公平性。载于公平、问责与透明度会议论文集。ACM，纽约州纽约，美国，39–48页。DOI：http://dx.doi.org/10.1145/3287560.3287567[35] Samir Passi 和 Steven J. Jackson。2018。数据科学中的信任。*ACM*会议论文集

人机交互 2，计算机支持的协同工作（112018），1–28页。 DOI：http://dx.doi.org/10.1145/3274405[36] Samir Passi 和 Phoebe Sengers。2020。让数据科学系统发挥作用。*Big Data & Society* 7，2（2020年7月），205395172093960。DOI：http://dx.doi.org/10.1177/2053951720939605[37] Kathleen H. Pine 和 Max Liboiron。2015年。测量与行动的政治。载于第*33*届 *ACM* 人机交互系统中的人为因素会议论文集。ACM，纽约州纽约，美国，3147–3156页。DOI：http://dx.doi.org/10.1145/2702123.2702298[38] Bogdana Rakova、Jingying Yang、Henriette Cramer 和 Rumman Chowdhury。2021。负责任人工智能与现实相遇。*ACM*人机交互论文集 5，计算机支持的协同工作1（2021年4月），1–23页。DOI：http://dx.doi.org/10.1145/3449081[39] Nithya Sambasivan、Shivani Kapania、Hannah Highfill、Diana Akrong、Praveen Paritosh 和 Lora M Aroyo。2021。“每个人都想做模型工作，而不是数据工作”：高风险人工智能中的数据级联。载于*2021*年 *CHI* 人机交互系统中的人为因素会议论文集。ACM，纽约州纽约，美国，1–15页。DOI：http://dx.doi.org/10.1145/3411764.3445518[40] Claudio Sarra。2020。将辩证法引入机器：通过更深入地理解设计中的可争议性来防范自动化决策。*Global Jurist* 20，3（2020年10月）。DOI：http://dx.doi.org/10.1515/gj-2020-0003[41] Sebastian Schelter、Yuxuan He、Jatin Khilnani 和 Julia Stoyanovich。2019年。FairPrep：在公平性增强干预研究中推动数据成为一等公民。（2019年11月）。[42] Jakob Schoeffer、Maria De-Arteaga 和 Niklas Kuehl。2022。关于解释、公平性感知与决策之间关系的研究。（4 2022）。[43] Andrew D Selbst 和 Julia Powles。2017。实质性信息与解释权。*InternationalData Privacy Law* 7, 4（11 2017），233–242。DOI：http://dx.doi.org/10.1093/idpl/ipx022[44] Divy Thakkar、Azra Ismail、Pratyush Kumar、Alex Hanna、Nithya Sambasivan 和 Neha Kumar。2022。机器学习数据何时才算优质？：公共卫生数据化中的价值评估。载于*CHI* 人机交互系统中的人为因素会议。ACM，纽约，纽约州，美国，1–16。DOI：http://dx.doi.org/10.1145/3491102.3501868[45] J. W. Thibaut 和 L. Walker。1975。《程序正义：心理学分析》。*L. Erlbaum Associates*，希尔斯代尔。（1975）。

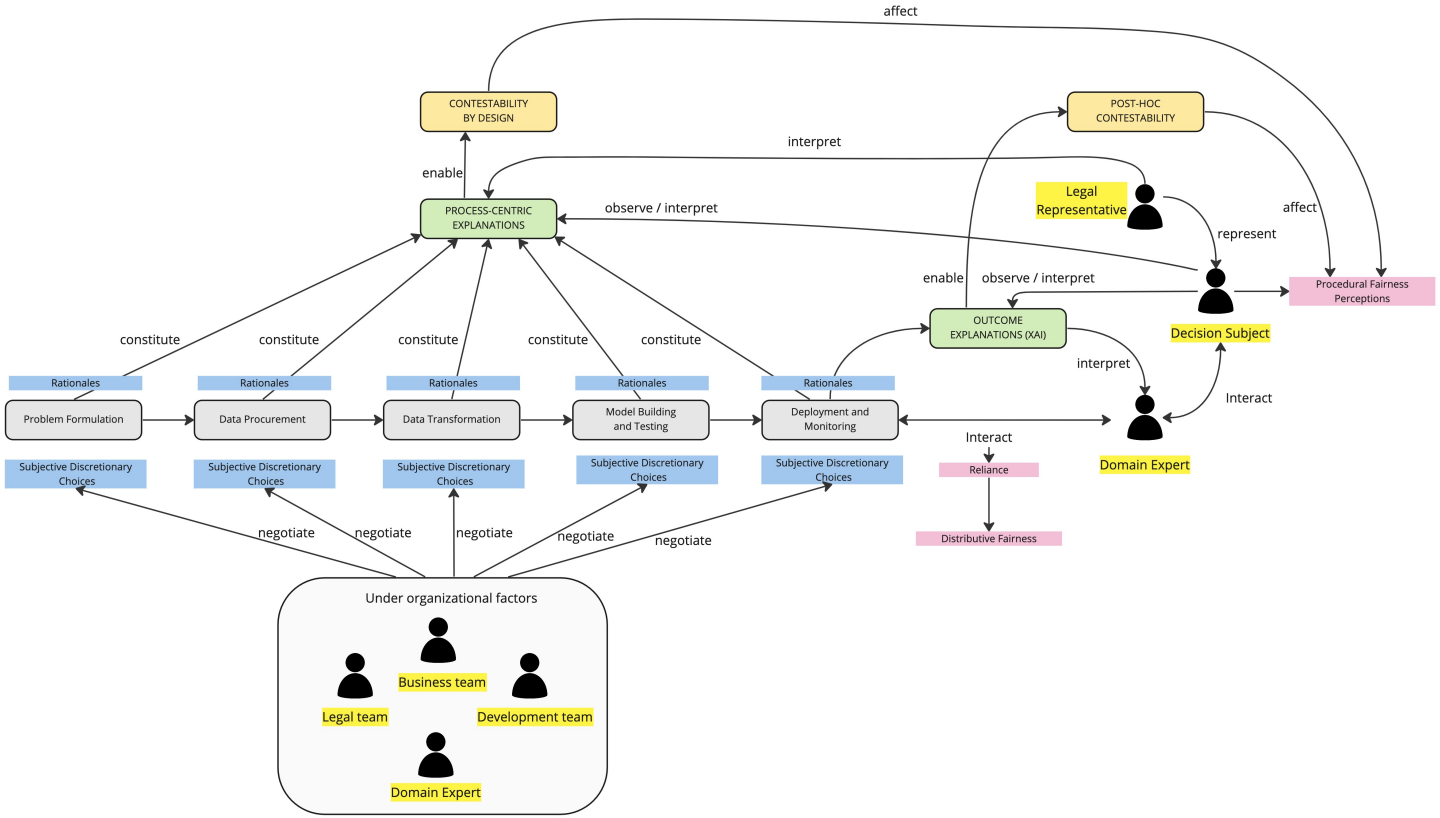

图 1： 概念框架概述：该框架将以过程为中心的解释定位为设计中的可争议性的必要要素，而设计中的可争议性进而影响决策主体的公平性感知。粉色部分标示 Schoeffer 等人定义的概念 [42]。灰色表示机器学习流程的不同阶段，绿色表示解释性要素，橙色表示可争议性要素，黄色表示不同的利益相关者，蓝色表示自由裁量选择及其理由。