算法批评的伦理边界
Ethical Boundaries of Algorithmic Criticism

万明泊（**Wan Mingbo**）

内容摘要：在生成式人工智能时代，文学研究面临着伦理与方法论挑战。斯蒂芬·拉姆齐的“算法批评”理论及其“变形阐释”理念，在此语境下需要被重新审视。算法的不透明性不仅构成技术难题，更引发阐释学危机，悬置了人类阐释的主体性。“可解释性人工智能”的技术解释无法替代人文阐释，而算法偏见的本质是内嵌于数据与模型中的价值负载，这使得数据策展与模型构建本身成为一种先于代码的隐蔽阐释行为。对此，亟须构建一种人机协同的阐释学实践，将算法视为具有自身视域的对话伙伴，而非权威答案的提供者。
关键词：算法批评；人工智能；伦理；阐释
作者简介：万明泊，南开大学文学院博士后，研究方向为阐释学研究。本文为中国博士后科学基金资助项目【项目批号：2025M773850】阶段性成果。

Title: Ethical Boundaries of Algorithmic Criticism

Abstract: In the era of generative artificial intelligence, literary studies face profound ethical and methodological challenges. Stephen Ramsay’s theory of “algorithmic criticism” and its concept of “deformance” require reexamination in this context. The opacity of algorithms not only poses technical difficulties but also triggers a hermeneutic crisis, marginalizing human agential subjectivity in interpretation. Technical explanations from “explainable AI” cannot substitute for humanistic interpretation, while the nature of algorithmic bias lies in the value-ladenness of data and models, rendering data curation and model construction a form of concealed interpretive act that precedes coding. In response, there is an urgent need to construct a human-machine collaborative hermeneutic practice, regarding algorithms as dialogic partners with their own horizons rather than as providers of authoritative answers.

Keywords: algorithmic criticism; artificial intelligence; ethics; interpretation

Author: Wan Mingbo is Postdoctoral Researcher at the College of Literature, Nankai University (Tianjin 300071, China). His academic research focuses on hermeneutics (Email: wanmingbo@nankai.edu.cn).

随着大型语言模型（LLMs）逐步渗透进知识生产的各个角落，文学研

究正遭受着一场前所未有的范式冲击。这不单单是研究工具或方法的更新换代，它直接触动了文学阐释的根基。在这块原本由人类主导、充满了价值判断与伦理考量的意义生成领域，当算法具备了模仿、续写乃至创造复杂风格文本的能力，那些根本性的文学伦理议题就变得无可回避：在技术的介入下，阐释的边界该划在哪里？谁才拥有生产意义的真正权力？在这场人机互动中，人文价值又会面临怎样的重塑、扭曲或是捍卫？

这些问题之所以迫切，是因为自然语言处理（NLP）技术从诞生起，就从来不是什么中立的工具。霍维（Dirk Hovy）与斯普劳特（Shannon L. Spruit）早就指出，语言其实是“人类行为的代理，以及个体特征的强烈信号”（592）。这意味着，任何处理语言的技术，处理的其实都是一套承载着社会关系、权力结构与意识形态的复杂系统。在早期的数字人文实践中，计算机也许还能被看作是整理文献、编制索引的仆从，但生成式 AI 的兴起，让它变成了一个能主动介入意义生成的行动者。这种介入正是本研究最为关心的：它把算法从外部的分析工具，变成了一个内在于阐释过程、充满伦理风险与可能性的准主体。

一、算法的介入：从阐释辅助到伦理挑战

斯蒂芬·拉姆齐（Stephen Ramsay）的算法批评理论，可以说是对上述挑战最早，也最具理论自觉的回应。在其著作《阅读机器：走向算法批评》（*Reading Machines: Toward an Algorithmic Criticism*, 2011）中，拉姆齐敏锐地洞察到计算方法与文学批评方法之间的根本性断裂（fundamental disjunction）[1]，前者追求的是可验证、可量化的客观事实，而后者则致力于通过主观介入来开启和深化讨论。面对这一断裂，拉姆齐并未选择让文学研究科学化，而是革命性地提出，我们应当将计算的狭窄约束——那种对枚举、测量和验证的执着——转化为一种服务于人文阐释的、富有创造力的工具。[2]

为此，拉姆齐阐述了作为其核心方法论的“变形阐释”（deformance）。他认为，任何真正的批评行为在本质上都是一种变形，“致力于提出一种‘读法’的批评家，他提出的不是文本本身，而是一个新的文本，其中的数据已被转述、阐发、挑选、删节和转化”（16）。算法批评正是将这种变形操作以一种严格的、整体的方式（rigidly holistic manner）加以执行。通过对文本进行系统性的重组、扭曲或可视化，算法生成了一个全新的副文本（paratext），其目的“不是为了寻求事实，而是为了寻求模式”（17）。

若从伦理视角审视，拉姆齐的理论其实藏着一种潜能。它把阐释的重

1 参见 Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press, 2011, 8.

2 参见 Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press, 2011, 16.

点从“忠实再现作者原意”，转移到了“通过创造性建构生成新意义”。它鼓励批评家“将文化产物视为可被计算工具转化、重组与再造的对象”，最终目的不是为了获得确凿真理，而是为了“不断抵达更深入、更复杂的新问题”（85）。这在理论上捍卫了阐释的开放性，并许诺了一种更民主的意义生产方式：只要能开启有效对话，任何算法变形都可能成为合法的阐释路径。

尤其是在当前时代，AI 拥有处理海量多语言文献的能力，能够跨越语言上的障碍，去挖掘那些长期被主流叙事所遮蔽的史料。大型语言模型能够生成类似人类的文本和代码，这为处理海量的历史文献和构建新的叙事模型提供了工具。在数字人文的框架下，人工智能可以处理和传播相关的主题内容，从而可能揭示出传统研究方法难以触及的历史关联。例如，通过对海量非英语文献的深度学习与分析，AI 理论上可以协助学者还原被忽视的文明贡献，从而为文明互鉴提供坚实的数据支撑。

然而，这种理论上的解放潜能在实际操作中极易遭遇伦理困境。问题的核心在于早期理论家忽略了一个关键事实：算法及其依赖的数据，从来都不是一张白纸。正如大量工具批评研究揭示的，对技术“不加反思的采用，可能会损害研究结果的合理性和可复现性”（Herrmann 1），算法工具里早已深深嵌入了设计者的理论预设与价值偏见。[1] 这种预设往往更为隐蔽且强硬。当这一点与大型语言模型那些庞大且未经审视的训练数据结合时，伦理危机便随之爆发。这些模型从互联网上抓取文本进行学习，而这些文本本身就反映着社会既有的权力结构与偏见，结果必然是“复制这些伤害”（Weidinger 1）。霍维与斯普劳特的研究也指出，NLP 模型训练数据中普遍存在“人口统计学偏见”，比如过度代表“WEIRD”人群（西方的、受过教育的、工业化的、富裕的、民主的），而让其他群体处于边缘（593）。

布洛杰特（Su Lin Blodgett）等人的大规模调查则更为系统地揭示了这一问题的严重性，他们将算法偏见造成的伤害归纳为两大类：“分配性伤害”（Allocational harms），即系统不公平地分配资源或机会；以及“再现性伤害”（Representational harms），即系统以贬低、刻板或侮辱性的方式再现特定社会群体（2）。这些伤害在文学阐释领域尤为致命。正如聂珍钊指出，“尽管 AI 作家用算法和数据分析代替伦理选择提高了文学生成的效率，突出了文学生成的科学性，但是现阶段的 AI 文学仍然是伦理选择的产物”（“科学选择与 AI 文学” 16）。当一个自带偏见的算法对文本进行变形时，它生成的模式必然会放大并合法化这些偏见。比如，由于训练数据的偏差，用于检测仇恨言论的模型更倾向于将非裔美国人白话英语（AAE）标记为“有毒”（Xu 3），这种技术判断本身就构成了一种“语言不公正”（Craft 391）。

这直接把算法批评的伦理潜能给颠覆了。原本寄希望于借助变形操作来

1　参见 J. Bereniken Herrmann et al., “Tool Criticism in Practice: On Methods, Tools and Aims of Computational Literary Studies,” *DHQ: Digital Humanities Quarterly* 3 (2023): 2.

打破阐释霸权，并由此开启多元意义，但事与愿违，它可能逐渐演变成对偏见的加深复读。它不仅没能实现阐释的民主化，反倒披着看似客观的技术外衣，固化了既有的社会等级与文化偏见，给阐释边界和意义生产权套上了更隐蔽的枷锁。因此，当代关于算法批评的任何讨论，都必须直面这一伦理困境：我们如何在利用算法开启阐释可能性的同时，抵制其内嵌的偏见，并朝向一种真正的阐释正义？这些才是我们需要深入探讨的核心议题。

二、“黑箱”的伦理困境与可解释性的限度

算法在对文本进行变形操作时，往往会无意识地复制甚至放大社会的既有偏见。这个问题的症结，归根结底在于大型语言模型的一个本质特性——不透明性（opacity），也就是人们常说的“黑箱”难题。如果单从技术上看，黑箱无非意味着我们没法用人脑能理解的方式，去一步步倒推模型内部那亿万个参数是如何互动并最终产出某个结果的。可一旦切换到文学伦理学的视角，这个黑箱就不再只是个技术障碍了，它引发的是一场阐释学危机。

我们之所以认为文学批评具有合法性，根本上是因为它的论证过程是可追溯、可辩驳的。一位合格的批评家，总能把从文本证据到理论结论的推导链条，并摊开给读者。但当一个黑箱模型对文本下判断时——比如把某种方言定性为低俗，或者把某个角色的行为归类为刻板印象——它抛出来的只是一个结论，却没有论证过程。这种“表征的暴力”之所以可怕，在于它披着技术权威的外衣，直接没收了阐释活动本该有的对话与思辨空间。正如卡西尔扎德（Atoosa Kasirzadeh）所言，在涉及刑事司法等高风险决策时，这种不透明性意味着“人类难以找到充分理由去理解，为何在特定情境下会得出这样一个算法结果”（2）。当这种不透明性被移植到同样充满价值判断的文学阐释领域，结果就是阐释权力的悄然让渡：权力从人类批评家手中，滑落到了一个我们根本无法质询其动机的算法系统里。

这一点在布洛杰特等人对 NLP 领域偏见研究的批评中得到了佐证。他们发现，很多研究的动机显得“模糊、不一致，且缺乏规范性推理”（1）。这其实恰恰勾勒出了黑箱阐释的尴尬：它或许能识别出某种模式（比如非裔美国人白话英语常与负面情感挂钩），却说不出这种关联为何有害、对谁有害，以及其中的伦理逻辑何在。这本质上是一种缺乏伦理自觉的阅读，它产出的意义悬浮在半空，根本没法进入人类的价值辩论体系。

面对黑箱带来的伦理与信任危机，计算机科学领域的回应是发展“可解释性人工智能”（Explainable AI, XAI）。XAI 的目标是开发一系列技术方法，以使 AI 系统的决策过程对人类用户更加透明。其社会与伦理动机在于：“增强对基于预测的决策的社会接受度，在这些决策结果中建立信任，使算法对公众负责，并消除算法歧视与不公的来源”（Kasirzadeh 2）。从表面上看，这似乎为算法批评的伦理困境提供了一条出路。然而，从文学伦理学批评的角

度深入审视，XAI 所提供的“解释”与人文研究所追求的“阐释”之间，存在着一条巨大的鸿沟。XAI 提供的解释，多半是技术性、描述性的。比如，它可能会告诉你，模型之所以判定某段文本“有毒”，是因为训练数据里某个特定的词（比如某个特定群体的俚语）和“有毒”标签高度相关。

这种解释在技术层面是诚实的，但在阐释层面却是苍白的。它只回答了“是什么”（What），也就是什么技术因素导致了这个结果；却完全回避了“为什么”（Why），即这种关联本身为何值得警惕？它背后折射出怎样的社会权力结构与语言意识形态？[1] 而这后者，恰恰才是人文阐释的灵魂。正如布洛杰特等人倡导的，负责任的研究必须讲清楚“对‘偏见’是如何概念化的——即什么样的系统行为是有害的，以何种方式，伤害了谁，以及为什么”（5）。遗憾的是，XAI 的解释往往止步于对系统行为的技术描述，无法自动完成这一步规范性的、价值驱动的跨越。

所以，如果过度依赖 XAI，我们很可能会陷入一种新的伦理风险：把复杂的阐释问题降维成技术诊断问题。研究者可能会满足于修补那个导致偏见的技术环节（比如调整算法权重来去偏），却忘了反思产生这一偏见的深层社会文化土壤。这就是一种典型的“技术方案主义”（techno-solutionism）：打着可解释性的旗号，巧妙地绕过了真正棘手的伦理与政治议题，治标却从未治本。

反思 XAI 的局限性，最终把我们的目光引向了算法阐释的真正源头，就是训练数据，以及嵌入系统设计中的价值观。算法之所以会有偏见，最根本的原因在于它“忠实地反映了训练数据中那些不公正、有毒和压迫性的话语”（Weidinger 6）。这些数据从来不是世界的客观镜像，而是人类社会既有权力关系、历史偏见和文化冲突的沉积物。因此，收集和标注数据，这本身就是一种极其强势的、先于算法运行的阐释行为。正如有研究者从后殖民理论视角指出的，“价值观和权力才是这场讨论的核心”（Mohamed 2），AI 技术的应用很可能在延续甚至加剧历史上形成的权力模式。决定把哪些文本拉进训练集、把哪些踢出去，以及怎么给这些文本贴上“积极”“消极”“有毒”或“正常”的标签，这一系列策展（curation）动作，从一开始就给算法规定了看待世界的视域。

卡西尔扎德在其哲学框架中将此称为 AI 推理的“背景假设和价值负载”（Background assumptions and value-ladenness of AI inferences）（20）。她指出，AI 系统在“本体论、认识论和测量等多个层面”都充满了价值偏见。例如，当一个系统被设计用来预测犯罪风险时，设计者必须首先对“犯罪”进行操作化定义，而这个定义本身就充满了社会和政治的价值判断。这最终将我们带回了文学伦理学的核心议题：意义的生产权。在生成范式下，权力不仅体现在对文本的最终解读上，更体现在对算法“前见”的塑造上。科技公

1　参见 Justin T. Craft et al., “Language and Discrimination: Generating Meaning, Perceiving Identities, and Discriminating Outcomes,” *Annual Review of Linguistics* 6 (2020): 392.

司、数据工程师和标注工人，在他们选择数据、设计模型、制定标注指南时，就已经在很大程度上预设了算法看待世界的方式。与传统文本中可被读者重构的“隐含读者”不同，AI 模型中的“隐形读者”在建构之初便被技术手段深度编码并“刚性固化”，形成了一种难以撼动的强势在场（任洁 85）。他们的选择，无论是有意的还是无意的，都构成了对人类文化文本的一次大规模预阐释。而黑箱的存在，则巧妙地将这次预阐释中蕴含的权力关系与价值偏见遮蔽了起来。因此，算法批评的伦理学任务，必须深入到这一后台之中，去质询和揭示那些在代码运行之前，就已经被悄然做出的价值决断。

行文至此，算法批评的伦理边界问题已然浮现。前文的论述反复触及这一边界如何被数据的偏见所侵蚀，但对边界本身的正面勘定，仍有待展开。在算法批评的语境下，“伦理边界”并非一条旨在阻挡技术介入的僵硬红线，它更应被理解为一个动态的规范性框架，用以校准和引导人机协同的阐释实践。因为任何由算法生成的阐释材料，无论是模式的揭示抑或文本的生成，都不能再以黑箱的神秘面目示人。它必须在最大程度上敞开其技术生成的路径，清晰地交代其所依赖的“设计解释”与“数据解释”（Kasirzadeh 2）。这意味着，一个阐释结论不仅要呈现结果，更要回答它是如何被知道的，从而将算法的答案重新置于人类理性可审查的链条之上。

与此紧密关联的，是对伤害的可预见性与最小化。当算法的介入可能造成分配性或再现性的伤害时，伦理的边界便已临近。这要求每一次阐释，都应以对模型偏见的严格审查为前提。例如，研究已反复证明，由于训练数据中存在偏见，用于检测仇恨言论的模型，会不成比例地将少数族裔正常言论错误地标记为“有毒”或冒犯性。在这种情况下，若不加反思地运用此类模型去发现其文学中的情感模式，其行为本身就已经构成了一种再现性伤害，其学术合法性也便荡然无存。任何依赖此模型对相关文本的分析，都已然越过了伦理的边界。

所以，一个合乎伦理的算法阐释，还需体现出对语境整体性的尊重。这种尊重具有双重维度：其一，是文学文本自身的历史语境；其二，则是算法模型及其训练数据的语境。大多数 NLP 模型依赖的是“WEIRD”语料，其所信奉的往往是一种“同质化偏好所形成的标准语言意识形态”（Craft 390）。用这样一个带有强烈偏见的视域去解读一部非西方的文学作品，无异于一场预设了结论的对话。忽视算法自身的语境局限，将其输出普遍化为对文本的客观发现，不仅削平了阐释的复杂性，更是一种智识上的懒惰。因而当我们运用大型语言模型分析文本时，正需要这种伦理边界加以限定与引导。

三、走向人机协同的阐释学：构建负责任的算法批评实践

前文试图揭示一个核心困境：一方面，算法批评以其强大的变形能力，允诺了一种更为开放和多元的意义生产模式；另一方面，算法黑箱及其内嵌的

数据偏见，又可能使其沦为固化社会不公、削弱人类阐释主体性的工具。处在这样的一种困境下，我们既不能因噎废食，退回前数字时代的技术恐惧中，也不能天真地拥抱那种缺乏反思的技术乐观主义。我们需要做的，是跳出这组二元对立，去探索如何在实践中打磨出一套负责任的、以人类价值为锚点的算法批评方法论。其中的关键，在于要把人机关系从单向度的“工具 — 使用者”模式，重塑为一种充满反思、对话与批判张力的人机协同阐释关系。

构建负责任的算法批评，第一步就得把工具批评从一种事后的外部反思，内化为贯穿技术设计与应用全周期的前瞻性实践。这与数字人文领域倡导的“批判性技术实践”（Critical Technical Practice）不谋而合（Mohamed 14）。它要求研究者不能只做算法的用户，更要成为具备批判意识的建造者。这意味着，文学伦理学批评的考量必须前置到算法批评的每一个毛细血管里。在数据策展阶段，我们必须摒弃对大数据的盲目迷信，转而拥抱一种“小数据”的、讲究策展伦理的实践。在搭建文学分析专用的训练集时，必须清晰地把数据的身世——来源、构成、潜在偏见以及那些被挡在门外的声音——都记录在案。正如有学者强调的，负责任的研究得让“数据集中所代表的群体、样本和叙事，以及那些可能缺失的拼图，都能被看见”（Weidinger 3）。

在模型设计与微调环节，文学研究者需要从书斋里走出来，与计算机科学家联手，探索如何把人文价值纳入模型里。比如，沙纳汉（Murray Shanahan）与克拉克（Catherine Clarke）在评估模型创造力时，制造出“创造性对话”（Creative Dialogue）和“多声音生成”（Multi-Voice Generation）的交互策略（3），这本身就是把文学批评里讲究的对话性与多声部理论，内化进了算法的交互设计之中。

而论及界面与交互设计，工具的界面本身就是伦理的载体。一个负责任的工具，它的界面应该能诱发用户的反思与批判。试想，当屏幕上跳出一个分析结果时，能不能顺便把置信度标出来？能不能揭示一下它依赖了哪些关键数据特征？甚至直接给出链接，指向关于该数据偏见的批判性文献？这才是把可解释性从技术后台推向用户前台的伦理设计。

这种批判性技术实践，促使文学研究者不能再把技术当成一个黑箱，而是必须在一定程度上厘清它的脾气秉性，并积极介入到它的设计与改造中。这倒不是说每个文学学者都得去写代码，而是倡导一种深度合作、反思共进的模式。而在技术介入之外，更深层的变革发生在阐释观念上。在这种协同范式下，我们实际上正在从事一种聂珍钊所定义的“计算分析批评”。其核心要义，在于将机器强大的“计算能力”转化为批评者敏锐的“洞察能力”（“AI 阅读与文学的计算分析批评” 31）。负责任的算法批评，必须自觉地把人机交互构建成一个伽达默尔意义上的阐释学循环。也就是，不将 AI 当成有问必答的百科全书，而要把它看作一个有着独特视域（虽然这视域也是被建构的）的对话伙伴。

AI 的视域由其训练数据和算法结构所决定，它充满了偏见，但同时也蕴含着人类个体无法企及的广阔模式。而人类批评家的视域，则由其深厚的理论素养、历史知识和价值关怀所构成。真正的阐释性洞见，正是在这两个视域的碰撞、对话与融合（Fusion of Horizons）中产生的。这种对话的本质是批判性的。人类学者必须时刻对 AI 的输出保持一种阐释学的怀疑（hermeneutic suspicion）。当 AI 生成一个看似新颖的模式或结论时，我们的首要任务不是接受它，而是质询它：这个模式是源于文本的内在结构，还是仅仅是训练数据中的统计偶然？这个结论反映了文本的复杂性，还是将文本强行纳入了模型所偏爱的某种简化框架？如果我换一种提问方式，引入一个不同的理论视角，模型的回答会发生怎样的变化？

沙纳汉与克拉克的研究生动地展示了这一点。在他们的实验中，人类用户扮演评论家或导师的角色，通过不断地追问、建议和风格引导，与 ChatGPT 进行反复的迭代，最终共同生成了具有相当文学复杂性的文本片段。[1] 在这个过程中，人类的价值判断、审美偏好和理论框架，始终是对话的主导力量。AI 的生成能力被用作一个强大的灵感激发器，但其产出始终处于人类批评家的审视与引导之下。

有鉴于此，我们可以将算法批评与提示工程（prompt engineering）、微调（fine-tuning）等计算技术相结合的介入性实践。尽管大型语言模型的整体倾向是西方中心的，但其庞大的训练数据中，依然包含了海量的关于非西方文明的知识。问题的关键在于，这些知识在通常情况下处于休眠状态。我们可以通过精心设计的提示，来主动激活这些休眠的知识。比如，有意识地引导模型扮演非西方的历史主体，迫使模型跳出其默认的叙事框架，去主动链接两个不同知识簇。研究表明，通过赋予模型一个具体的角色，可以有效地引导其表达特定的政治或意识形态立场。[2]

当前主流的 RLHF 范式，其目标是让模型与人类的偏好对齐，这是一种追求共识的努力。然而，一个富有活力的阐释共同体，也需要不同的声音与视角的碰撞。因此，我们借助扮演不同意识形态角色的 AI 代理，让它们在面对同一个问题时，给出基于不同世界观的回答。此路径，就是将大型语言模型不再仅仅视为一个文本生成工具，而是将其视为一个封装了庞大文化无意识的对象。研究者通过设计一系列有针对性的提问或情境，去发掘模型内隐藏的价值假设。这种方法在 NLP 偏见研究中已有广泛应用，例如，布洛杰特通过测试“医生”一词与男性代词的关联强度，对比“护士”与女性代词的关联强度，来揭示模型中的性别职业偏见（2）。在文学研究中，这种方法

1　参见 Murray Shanahan and Catherine Clarke, “Evaluating Large Language Model Creativity from a Literary Perspective,” *arXiv preprint arXiv*: 2312.03746 (2023).

2　参见 Elena Musi et al., “Toward Reasonable Parrots: Why Large Language Models Should Argue with Us by Design,” *arXiv preprint arXiv*: 2505.05298 (2025).

可以变得更为深入。比如我们对文学叙事进行探查，让大型语言模型去重写多部俄罗斯经典的结尾。通过分析模型生成的另类情节，我们可以勘探出该模型所学习到的当时俄国社会脚本是什么。它生成的或许是一个个平庸的故事，但这个故事本身，就成为反映我们文化集体想象的一面镜子。这种方法将算法批评从对单个文本的分析，拓展到了对文化原型与叙事惯例的批判性反思，使得算法黑箱本身，成为我们进行思想史研究的新对象。

通过将算法批评构建为一种批判性的、对话式的实践，我们最终得以回应那个核心的伦理焦虑——人类阐释主体性的维系。在负责任的算法批评框架中，人类与机器的分工是明确的，这种分工确保了人类作为意义的最终裁决者的地位。算法的角色是提问而非回答：即使算法以陈述句的形式输出，其在阐释学循环中的功能也应被视为一个提问。它提出的问题是："这里似乎存在一个模式，你认为它重要吗？它与你的理论框架有何关联？"人类的角色是赋予意义，面对算法生成的无数模式、关联和虚拟文本，只有人类批评家能够判断其价值。是人类学者将这些冰冷的数据，与文学史的脉络、哲学的思辨和社会关怀联系起来，从而赋予其以人文意义。

拉姆齐在其著作的结尾提出了一个愿景：他所期待的，是一种"对批评被天真地机械化不必担忧，对算法被过度使用也不必忧虑"的从容态度。[1] 这种从容，正源于对人机分工的自信。正如魏丁格等人在其风险分类报告中所反复强调的，对于许多 AI 伦理风险，最终的缓解措施都指向加强人类的监督和有意义的人类控制。[2] 在算法批评的语境下，这种控制权，正是阐释的最终裁决权。

我们所倡导的批判性技术实践与人机协同对话，其意义远不止于获得更高效的文学洞见。从根本上说，这是在回应开篇所提出的阐释正义的吁求。阐释正义，在此语境下意味着，我们所运用的阐释工具与方法，应致力于纠正既有的话语权力不平等。所谓批判性技术实践，正是对生产阐释的工具本身所行使的一种权力制衡。文学研究者深入技术后台，从数据源头和模型设计上挑战这种所谓的标准语意识形态。这不仅是方法论的完善，更是一场捍卫文化多样性的伦理行动。它回应了阐释正义的首要诉求：生产意义的工具，不应成为再生产"语言不公正"的帮凶。而人机协同对话，则为聆听与放大边缘声音提供了可能。一个不带偏见的算法，可以成为更好地把握他者的声音，揭示出被主流阐释所忽略的文本模式。更重要的是，通过质询一个有偏见的模型，我们恰恰能够反向地照亮那些在数据层面被边缘化的话语。通过坚守最终裁决权，批评家确保了对算法输出的价值判断，始终植根于对人的境遇的同情和对社会公正的追求。

1　参见 Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*, Urbana: University of Illinois Press, 2011, 85.

2　参见 Laura Weidinger et al., "Ethical and Social Risks of Harm from Language Models," arXiv preprint arXiv: 2112. 04359 (2021).

算法批评非但不会导向人文价值的终结，反而可能促成一次对其核心使命的再确认。在一个信息过载、模式泛滥的时代，文学批评的核心任务，即在纷繁的表象中做出有价值的判断、构建出有意义的叙事，正变得比以往任何时候都更加重要。算法可以成为我们探索意义疆域的强大盟友，但定义这片疆域的边界、并最终裁定其中何为珍宝的权力，必须也只能掌握在人类自己手中。

结语

算法的介入，已不仅仅是研究方法的革新，更是一场触及意义生产权力、阐释边界和人文价值核心的伦理事件。面对这一现实，文学研究既不能退守于对技术的人文主义式拒斥，也不能陷入对算法能力的非批判性崇拜。聂珍钊指出，“建构 AI 文学理论将从计算分析和计算批评开始”（“科学选择与 AI 文学” 15）。我们必须构建一种更具反思性、对话性和责任感的人机协同阐释学，以确保技术的发展最终服务于深化而非削弱我们的人文理解。

一方面，我们要重申拉姆齐理论的开创性价值。他提出的“变形阐释”，把计算的死板约束转化成了开启文本多元可能性的创造力，许诺了一种更民主的意义生产模式。它挑战了传统阐释对作者原意的执念，把重心转到了通过建构来生成新问题的动态过程上。但另一方面，我们也得清醒地看到这一理论在当下面临的伦理窘境。大型语言模型作为黑箱的不透明，加上训练数据里洗不掉的社会偏见，让原本旨在解放的变形操作，往往容易形成对偏见的加深复读。学者的研究已经拉响了警报：算法对特定社会群体语言模式的系统性误读与贬抑，已经构成了事实上的“语言不公正”（Craft 391）。这种打着技术中立旗号的表征暴力，对阐释正义构成了威胁。

面对这个算法所带来的困境，我们不能指望“可解释性 AI”（XAI）能一劳永逸。虽然 XAI 在技术上让模型稍微透明了一点，但它给出的技术解释和人文研究追求的价值阐释，根本无法等量齐观。如果不假思索地接受 XAI，很可能把伦理与政治问题，降维成修修补补的技术诊断问题，从而遮蔽了偏见产生的社会根源——那些在数据收集、标注与模型设计之初，就已经悄悄做出的“价值负载的决断”（Kasirzadeh 20）。

因此，出路在于一种走向人机协同的、批判性的阐释学实践。其核心是把人机关系从简单的“工具－使用者”模式，重塑为伽达默尔式的阐释学循环。在这个循环里，AI 不是提供标准答案的权威，而是那个拥有独特视域、能挑战我们固有成见的对话者。人类学者的主体性，恰恰是在这种持续的、由人类主导的批判性质询中立住的。研究表明，最有效的人机协同，就是人类扮演导师，通过不断追问、引导和修正，去激发和塑造 AI 的生成能力。[1]

这最终重新定义了算法时代的阐释责任。在未来，一个负责任的算法批

1　参见 Murray Shanahan and Catherine Clarke, “Evaluating Large Language Model Creativity from a Literary Perspective,” *arXiv preprint arXiv*: 2312.03746 (2023).

评实践者，得修炼出三重技法：首先是技术素养的伦理化，要懂技术，更要能批判性地介入算法的设计与应用，把工具批评变成贯穿始终的实践。[1] 其次是阐释过程的对话化，要有技巧地跟 AI 进行批判性对话，设计出既能激发模型潜力、又能暴露其局限性的阐释实验。最后是意义判断的人本化，在算法生成的无数可能性中，坚定人类学者作为最终意义裁决者的地位，用深厚的人文关怀和理论洞见，做出有分量的判断。

回到拉姆齐的愿景，他希望有一天算法批评能像“基于图书馆的批评”一样，成为一个无需多言的术语。[2]在生成式AI时代，这个愿景有了新的注脚。AI正在变成我们这个时代最庞大、最复杂的图书馆，它不仅储藏信息，还能生成话语。学会如何在这座图书馆里进行有伦理自觉的阅读、对话与创造，是我们这一代人文社科学者逃不掉的使命。因为这不仅仅是工具的扩张，更是一场本体论层面的“科学转向”。正如聂珍钊所言，“在可以预见的未来，传统的人文必然转化为数字人文和智慧人文”（“人文研究的科学转向” 568）。将文学理论与科学思维相融合，我们才能突破传统人文研究的困局。算法批评的未来，不在于让机器变得更像人，而在于通过与机器的互动，让我们对自己作为“人”的阐释责任，有多一分的自觉。这不仅是文学研究的未来，更是数字时代人文精神得以延续和发展的希望所在。

Works Cited

Blodgett, Su Lin et al. “Language (Technology) is Power: A Critical Survey of ‘Bias’ in NLP.” *arXiv preprint arXiv*: 2005.14050 (2020).

Craft, Justin T. et al. “Language and Discrimination: Generating Meaning, Perceiving Identities, and Discriminating Outcomes.” *Annual Review of Linguistics* 6 (2020): 389-407.

Herrmann, J. Berenike et al. “Tool Criticism in Practice: On Methods, Tools and Aims of Computational Literary Studies.” *DHQ: Digital Humanities Quarterly* 3 (2023): 1-30.

Hovy, Dirk and Shannon L. Spruit. “The Social Impact of Natural Language Processing.” *The 54th Annual Meeting of the Association for Computational Linguistics Proceedings of the Conference* 2 (2016): 591-598.

Kasirzadeh, Atoosa. “Reasons, Values, Stakeholders: A Philosophical Framework for Explainable Artificial Intelligence.” *arXiv preprint arXiv*: 2103.00752 (2021).

Mohamed, Shakir, Marie-Therese Png and William Isaac. “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence.” *Philosophy & Technology* 4 (2020): 1-28.

Musi, Elena et al. “Toward Reasonable Parrots: Why Large Language Models Should Argue with

1 参见 Shakir Mohamed, Marie-Therese Png and William Isaac, “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence,” *Philosophy & Technology* 4 (2020): 18.

2 参见 Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism*, Urbana: University of Illinois Press, 2011, 81.

Us by Design." *arXiv preprint arXiv:* 2505.05298 (2025).

聂珍钊："AI 阅读与文学的计算分析批评"，《外国文学研究》5（2025）：22-32。

[Nie Zhenzhao. "AI Reading and Computational Analytical Criticism of Literature." *Foreign Literature Studies* 5 (2025): 8-17.]

——："科学选择与 AI 文学"，《外国文学研究》3（2024）：8-17。

[—. "Scientific Selection and AI Literature." *Foreign Literature Studies* 3 (2024): 8-17.]

——："人文研究的科学转向"，《文学跨学科研究》4（2022）：563-568。

[—. "The Scientific Turn of Humanities Studies." *Interdisciplinary Studies of Literature* 4 (2022): 563-568.]

Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. Urbana: U of Illinois P, 2011.

任洁："伦理身份与 AI 时代的文学伦理学批评"，《外国文学研究》5（2025）：77-88。

[Ren Jie. "Ethical Identity and Ethical Literary Criticism in the Age of Artificial Intelligence." *Foreign Literature Studies* 5 (2025): 77-88.]

Shanahan, Murray and Catherine Clarke. "Evaluating Large Language Model Creativity from a Literary Perspective." *arXiv preprint arXiv*: 2312.03746 (2023).

Weidinger, Laura et al. "Ethical and Social Risks of Harm from Language Models." *arXiv preprint arXiv*: 2112. 04359 (2021).

Weidinger, Laura et al. "Taxonomy of Risks Posed by Language Models." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022): 214-229.

Xu, Albert et al. "Detoxifying Language Models Risks Marginalizing Minority Voices." *arXiv preprint arXiv*: 2104. 06390 (2021).