人工智能评估、审查与审计的历史方法

（Historical Methods for AI Evaluations, Assessments, and Audits）

Juana Catalina Becerra Sandoval
IBM Research
Yorktown Heights, USA
juana.becerra.sandoval@ibm.com

Felicia S. Jing
IBM Research
Yorktown Heights, USA
felicia.jing@ibm.com

摘要

本文提出将历史分析作为一项关键方法，纳入人工智能评估、审查与审计（AI evaluations, assessments, and audits）体系之中。尽管 FAccT 会议的学术传统已逐步认识到历史洞见的重要性，历史分析本身尚未以系统化、程序化的方式被正式纳入评估实践之中。我们将这一缺失置于人工智能审计（AI auditing）产业日益增长的政治与经济背景下考察，该产业的发展推动了评估方法的日益收窄，从而限制了可发现问题的范围。

本文借助我们此前在一项针对 AI 虚拟代理系统的影响评估中设计并实施历史分析的经验，提出一种可行的整合框架，旨在将历史分析与其他评估方法并行使用。我们进一步反思历史方法在人工智能评估中的适用性与价值，强调历史调查可用于识别更广泛的风险与影响，从而扩展评估协议中的“危害发现”（harm discovery）阶段。此外，我们还指出，历史分析所提供的洞见能够为评估发现提供有力支撑，并帮助将其置于恰当的社会、文化、政治与经济语境中加以呈现。

关键词

历史（history）、审计（audits）、评估（assessments）、评价（evaluation）、人工智能（artificial intelligence）、虚拟代理（virtual agents）

1 引言

人文学科、社会科学以及以人为中心的研究视角已被广泛认为是解决人工智能伦理、公平性与责任问题的关键要素 [74, 83, 89, 100, 113, 121, 143]。为评估人工智能系统所带来的风险与社会影响，相关研究努力发展出一系列框架与技术，强调应拓展社会技术评价（sociotechnical evaluation）方法的范畴，纳入包括访谈、民族志观察、参与式互动等在内的手段，以增强评估的情境敏感性 [16]，并借

助跨学科协作研究的严谨方法论基础 [82]。这些呼吁共同表明，识别与预测人工智能系统在过去、现在与未来的社会影响，是一个横跨多领域的系统性问题（transversal problem），需要超越技术本身，深入考虑技术的创建、整合、（再）利用、废弃与淘汰等过程所涉及的社会、文化、经济与政治语境 [109, 159]。

本文旨在回应并贡献于这一呼吁，提出将**历史分析（historical analysis）**作为一项关键的方法论组件，纳入人工智能评估、审查与审计（AI evaluation, assessment, and auditing）协议体系中。所谓历史分析，是指一类以“过去”为重要知识来源与研究对象的方法实践。在人工智能评估的语境中，历史分析意在通过借鉴过去的经验教训，理解当下、筹划未来。与其他情境性与质性分析方法相似，历史分析致力于揭示技术、社会、政治与文化之间的复杂关系。我们认为，其独特价值在于：它强调档案资料的重要性，并关注当代人工智能系统的深层历史根源。

在接下来的章节中，我们将历史分析引入当前有关人工智能审计（AI auditing）方法论的批判性研究文献之中 [26, 94, 111, 114, 147, 154]，并将其定位为一种可被整合进评估流程中的社会技术评估方法。我们认为，它可在多个层面发挥作用：拓展“危害发现”（harms discovery）阶段的广度；为研究问题的提出、测试设计与评估参与提供理论基础；并用于支撑与传播评估结论。虽然 FAccT 会议长期以来已认可历史视角对理解权力关系如何介入技术实践的重要性 [76]，并指出人工智能系统如何加剧既有社会不平等问题 [2, 10, 15]，但我们进一步主张，应以更加系统性、程序化的方式将历史分析正式引入人工智能评估实践。为此，我们强调应充分利用历史材料（如档案记录、口述历史、二手历史文献），采用历史研究方法（如地方史研究或“长时段分析”[longue-durée]），以及培养历史思维方式——即认识到过去如何持续塑造当下 [145]。

为了论证历史分析在人工智能评估中的重要性，**第二节**将从回顾 FAccT 会议以及相关学术界、公民社会与政策领域中既有的历史传统入手。通过梳理这些社群赋予历史思维的独特价值，我们认为，将历史作为批判性社会技术系统评估方法的做法，已有坚实的前例和严密的理论基础可依循。

接着在**第三节**中，我们将检视过去十年来人工智能评估、审查与审计（AI evaluation, assessment, and auditing）领域的重大变迁，追踪其从最初倡导公共监督的诉求，逐步演变为一个以审计产业为中心的局面。我们之所以展开这项考察，是为了更清晰地解释当前评估实践中所偏好的激励机制如何倾向于采用其他方法论 [31, 86, 141, 143]，并借此说明历史方法在当前多样化评估实践中为何仍显稀缺。

随后在**第四节**，我们将提出一种具体方案，用于将历史分析整合进人工智能评估、审查与审计流程之中。为展示该方法的实际可行性，第**五节**将介绍一个为期 16 周、涵盖多个学科的 AI 虚拟代理系统影响评估案例 [11]，该案例由本文两位作者共同设计并实施历史分析。通过这一案例，我们旨在展示历史方法所能带来的多重价值，尤其是在识别更广泛潜在风险与影响方面的作用，以及其如何为其他方法的评估结果提供佐证。

最后在**第六节**，我们将为未来有意将历史分析纳入评估体系的研究人员提供操作性指导，并反思在这一过程中所面临的机遇与挑战。

2 FAccT 中的历史研究传统

在 FAccT（公平性、责任性与透明度会议）相关学术研究中，历史分析与历史洞见已经被广泛采纳，并逐步构建出可称之为“历史研究传统”的学术谱系。在本节中，我们将简要概述并梳理历史在 FAccT 研究社群中被引入社会技术（sociotechnical）研究的多种方式。具体而言，历史被作为**参考框架、研究方法乃至实践路径**，被运用于以下研究方向：技术产业中的权力动态分析、影响人工智能系统治理的政治变迁研究、调节技术生产的文化与认识论规范探索，以及人工智能系统部署过程中所加剧的结构性不公现象分析，等等。

在概述这些历史研究路径的同时，我们也为本文的核心论点奠定基础——即将历史分析作为一项评估方法，正式纳入人工智能评估、审查与审计（AI evaluations, assessments, and audits）体系的必要性。我们将在已有研究的基础上进一步展开论证，说明这种整合不仅有其理论前提，也具备实践可行性。

2.0.1 将人工智能系统的设计、开发与实施置于历史语境之中

FAccT 社群运用历史分析的主要方式之一，是将人工智能系统的**发展过程、治理机制及其社会影响**置于历史语境中加以考察。此类研究包括对塑造当前技术系统评估与监管方式的历史转型的探讨，例如反歧视法律的兴起 [17, 110, 151, 152]、数据隐私政策的发展 [85]，以及人工智能监管框架的形成 [58]。

此外，还有一系列研究聚焦于**导致人工智能系统输入与输出不平等的历史性与结构性条件**，例如全球南北方之间的地缘政治不平等 [106]、种族歧视与隔离的历史 [13, 59, 128]，以及长期存在的技术官僚权力失衡 [150]。这些因素共同构成了影响人工智能系统开发与部署方式的重要背景。

在具体技术层面，FAccT 的研究者们还对若干人工智能子领域中的**认识规范（epistemic norms）**、**本体论假设（ontological assumptions）**与科学主张的历史渊源进行了梳理与提炼。这些领域包括计算机视觉系统 [46]、情绪识别系统与心理测量技术 [34, 71, 88, 134]、大型语言模型驱动的聊天机器人 [133]、生成式艺术 [45]、集中式开放代码实践 [28]、基准测试方法 [102]以及数据集构建策略 [130, 146, 157]等。

这些研究共同说明，人工智能系统并非在真空中被设计与部署，而是深深植根于历史制度与文化结构之中，理解这些背景对于开展负责任的人工智能评估具有关键意义。

2.0.2 借助其他领域的历史案例对人工智能干预进行问题化、阐释与启发

FAccT 社群运用历史的另一种方式，是通过跨领域的**比较分析（comparative analysis）**，考察缓解机制与治理干预在其他技术性与非技术性领域的历史演变，从而反思与丰富人工智能相关的干预策略。

例如，Hutchinson 与 Mitchell [67]，以及 Fröhlich 与 Williamson [42]，分析了“公平性”（fairness）这一概念在教育、招聘与保险等领域的历史应用，并将这些历史路径与当前机器学习（ML）研究中对公平性的处理方式进行对比，指出其中的历史延续性与可改进空间。

Chasalow 与 Levy [25] 也进行了类似的比较研究，但其聚焦于“代表性”（representativeness）概念，借助统计学与政治史中的典型案例，探讨其如何影响当代 AI 系统的评估标准。

Cooper 与 Vidan [29] 则深入研究了 20 世纪 70 至 80 年代互联网早期设计过程中的“征求意见书”（Request for Comments, RFC）制度，通过这些文献挖掘“问责制”（accountability）概念在不同阶段的演变与兴起。

最后，也是一个尤为相关的例子：Metcalf 等人 [94] 考察了**环境、人权与金融影响评估**的历史演进，从中提炼出关于“影响”这一概念的社会建构及其问责机制的关键经验，并将这些经验迁移应用于算法影响评估（algorithmic impact assessments）的制定过程中。

这些研究说明，从其他领域的历史实践中汲取经验，不仅有助于揭示人工智能治理中可能忽视的问题，还能为构建更具回应性的评估机制提供有力理论支撑。

2.0.3 从历史学科方法中汲取经验以指导人工智能相关实践

历史在 FAccT 研究中被转化为实际方法的第三种路径，是通过借鉴历史学科内部的方法论传统，提出一系列新的技术性、伦理性、分析性以及教育性的方法。这些方法不仅源于对历史的内容理解，更源自对**历史实践本身作为方法的反思与移植**。

例如，一些学者借鉴**档案管理实践**（archival practices）来改进人工智能系统中的数据收集与分类流程 [14, 68]；也有研究者受到**博物馆策展与展览实践**的启发，尝试将其应用于科技社群与使用群体之间的互动设计中，以提升包容性、多样性与可及性 [65]。

此外，许多研究者还主张引入源自**科学史与相关批判性学科**的理论与概念，以深化我们对人工智能系统影响的理解。这些理论与概念包括但不限于：

情境知识（situated knowledges）

具象化理论（figurations）

衍射方法（diffraction）

批判性虚构/推测（critical fabulation/speculation）

这些方法为我们提供了一种视角，帮助揭示人工智能系统背后的社会构造、文化预设与权力机制，从而避免将 AI 视为“中立”或“技术性纯粹”的产物 [49, 77, 79, 80, 96, 103, 121]。

2.1 历史的作用与价值

上文所述的 FAccT 中历史研究的广泛应用，表明该社群对“历史”这一维度已建立起深厚的关注与互动，也凸显出历史在回应其核心关切与价值取向中的关键地位。事实上，这些研究已形成了一个典范性的知识体系，展示了如何有效运用历史以达成以下目标：

深入理解技术的中介机制：包括底层科学理论、认识美德（epistemic virtues）、文化规范、社会想象、政治愿景与权力动态等，这些要素共同影响着技术的设计、开发与整合过程；

解释不平等的历史根源：揭示社会不平等的起源与结构性根基，为预测人工智能系统可能加剧这些不平等、并导致不均衡风险与危害提供前提条件；

从历史中提炼经验与预警机制：总结历史中的积极经验，警惕未来人工智能系统在设计、开发、（再）利用与治理中重蹈既往错误的覆辙。

在 FAccT 社群之外，**历史思维的实用性、历史类比的解释力**以及历史学者在处理技术相关问题上的专业能力，也逐渐受到关注与重视。早在 1970 至 1980 年代，关于**技术评估（Technology Assessments）**的早期方法论讨论就已提出，科学史与技术史研究者，以及科学与技术研究（Science and Technology Studies, STS）学者，理应参与到技术的监管性与应用性评估工作中 [55]。

近年的相关讨论同样延续了这一思路。例如，在**技术史学会（Society for the History of Technology, SHOT）**以及由亚利桑那州立大学“科学、政策与结果联盟（Consortium for Science, Policy, and Outcomes, CSPO）”组织的“科技政策：深入思考与实践建构（Technology Policy – Think Deeply and Build Things）”工作坊上，与会者都试图总结历史学者在技术评估与治理中可提供的洞见。特别是，那些“受过系统训练、能够分析复杂制度、经济与文化动态演变过程的历史学家” [30]，被认为在当前技术治理体系中具备独特而宝贵的视角。

与此同时，这些讨论也提出了对历史方法的局限性的反思，从而推动对**历史分析在方法论层面应用的细化与深化**。我们将在第六节对此进一步展开。

综上所述，FAccT 及其相关学术社群已经为历史方法的引入奠定了坚实的理论与实践基础，因此，历史方法在当前人工智能评估与审计（AI assessment and audit）协议中依然**缺位**，显得尤为值得关注。

在下一节中，我们将从评估模式与监管机制的演化过程出发，追溯过去十年间人工智能评估领域的重大转变，借此分析当前历史方法缺席的成因。我们也将反思这些转变背后的动因与方法论影响，指出当前评估模式在研究方法上的收缩趋势，并提出历史分析可作为拓展评估视野的有力路径，助力人工智能评估、审查与审计的进一步深化。

3 人工智能评估的简史

当前的人工智能评估与审计（AI assessment and auditing）形式[1]，**并非是评估方法自然演进或技术工具持续进步的结果**。相反，我们今天所理解的人工智能评估与审计，其实是**在特定社会与历史条件下形成的一套实践体系**——在其他条件下，它本可能呈现出完全不同的面貌。

毕竟，对于自动化、算法技术、机器学习方法以及数据驱动实践所带来的社会影响，批判性观点早已有之 [40, 90]。事实上，自人工智能技术诞生之初（甚至更早）起，关于其潜在风险与危害的担忧就已经开始流传。

那么，究竟是什么促成了我们今日所见的这一套特定的人工智能评估与审计生态？在接下来的若干小节中，我们将回顾过去十年来人工智能责任机制与治理模式中的几个重要阶段——从早期对公共监督的倡议，到 AI 审计产业的逐步形成。

在此基础上，我们将分析随之而来的具体评估方式与观察模式，并指出当人工智能评估被纳入产业结构时，其在**评估范围、速度与规模**上的标准化倾向所带来的影响。

3.0.1 公共监督

关于人工智能系统社会影响的研究，**在 2010 年代中后期开始获得更广泛的公众关注**。这一转变起始于一批学术研究者、活动家、记者与公民社会组织开始重新审视科技产业在以下方面所扮演的角色：加剧社会不平等、操纵公众政治意见、强化职场与国家监控机制，以及对环境退化的助长作用 [5, 12, 19, 21, 61, 62, 64, 98, 99, 101]。

与此同时，来自科技公司内部员工的法律行动与抗议活动也进一步提升了社会警觉性。特别是当一些组织如**科技工作者联盟（Tech Workers Coalition）**、

---

[1] 术语说明：本文所使用的“人工智能评估”一词，广义上涵盖了所有用于检视、识别和/或预判人工智能系统社会风险与危害的方式。然而，“评估（evaluation）”、“审查（assessment）”与“审计（audit）”三者之间在专业语境中仍存在重要区别：

审计（audit）通常被视为最为严格的审查形式，往往由第三方独立机构执行，且与技术开发过程保持一定距离 [16, 47]；

影响评估（impact assessment）则更适用于产品开发的早期阶段，其目标在于为设计过程提供前瞻性干预 [120]。

随着人工智能监管体系的逐步成型，以及评估标准的建立，这些术语间的区分变得尤为关键。尤其在涉及信息可获取性、利益冲突以及问责机制等问题时，不同评估方法所隐含的权力关系结构也成为必须严肃对待的议题 [115, 126]。

#TechWontBuildIt 与**#NoTechForICE** 相继发声，公开抗议科技公司与执法机关及武器制造商之间的合同合作关系 [123]，并揭露支撑许多人工智能系统背后的工作条件与劳动剥削问题 [104]，这一波监督浪潮逐渐浮出水面。

3.0.2 企业、监管与学术界的回应

在人工智能相关危害受到越来越多审视的背景下，来自企业、政府与学术界的应对努力开始逐步加强。事实上，早在 2018 年，科技行业的主要媒体——如 O'Reilly 与 TechCrunch——就曾将当年描述为"数据的清算日（Data's day of reckoning）"与"硅谷的清算之年（Silicon Valley's year of reckoning）"，指出科技产业正从一种无限制的自我授权模式，转向对外部责任的回应，或者至少是在姿态上进行"公开忏悔"。

这一时期，各方展开了多层面的响应举措，包括：

科技公司发布声明，宣布对"责任"与"科技伦理"的承诺 [38, 93]；

投资构建实践社群，如资助非营利组织、学术团体与会议平台 [49, 69, 95, 153]；

承诺在特定应用场景下逐步淘汰高风险技术 [3, 51, 52, 60]；

提高漏洞悬赏（bug bounty）、加强红队测试（red teaming），并推动公众参与，以揭示潜在风险与隐性影响 [75, 108, 138]；

更频繁地将人工智能审计（AI auditing）的建议与要求纳入法规制度 [47, 54, 141]。

这些措施共同构成了企业、监管机构与研究团体**对公众压力所作出的系统性回应**，也标志着人工智能技术治理逐步向制度化、程序化方向迈进。

3.0.3 人工智能审计产业的兴起

在上述背景下，**"识别与缓解人工智能系统潜在危害的压力，催生了对以公平性、问责性与透明性为核心的社会技术创新的紧迫需求"** [127]。这一趋势进一步推动了一个**人工智能评估与审计框架、方法与工具的市场**快速形成。

为应对日益增长的评估需求，多个组织开始开发相关服务，包括：

科技公司内部的专门团队；

各类知名咨询公司，如 PwC、EY、KPMG、德勤（Deloitte）、麦肯锡（McKinsey）与埃森哲（Accenture）；

获得风险投资支持的初创企业；

非营利组织，如 Apollo Research 与 METR；

以及一些专注于此领域的小型律所。

学界将这一趋势描述为**人工智能审计产业（AI auditing industry）的初步成型，并提出警示：这一产业的兴起可能导致一种新的政治经济格局，在其中，评估过程中所采用的方法、主导的行动者，以及被认可的专业知识类型，将呈现出日益收窄的倾向** [16, 141]。

在接下来的部分，本文将探讨当前人工智能评估与审计在形式上的主要特征，并进一步提出一个问题：**历史分析方法是否能够在此语境中拓宽我们对人工智能评估与审计可能性的理解与设想？**

3.1 对人工智能评估工作的影响

随着人工智能审计“市场”的兴起，**评估方法趋于统一化与标准化**的风险也随之上升。尤其在被要求**快速完成评估**的背景下，评估过程往往更倾向于：

使用**自动化方法**，

倚赖**通用的基准与标准**，

并优先采纳由大型成熟企业提供的服务。

在本节中，作者将进一步梳理此类发展对人工智能评估的**范围（scope）、速度（speed）、规模（scale）与评估场域（sites）所带来的具体影响，并就这种趋势可能对审计中的“可发现性空间”（terrain of discoverability）**所造成的限制，提出若干方法论层面的担忧。

3.1.1 范围（Scope）

有学者指出，当前人工智能评估与审计的工作**日益集中于“当下”的优先性问题**。换言之，这些评估主要关注技术的**当前能力**，以及其对**直接终端用户**可能造成的即时风险与危害。

这种聚焦模式导致评估视野往往被**限制在“实验室”内部** [43]，而忽略了更广泛的社会、历史、文化与政治语境。这种狭隘的评估范围使得评估工作的重心更倾向于**技术开发者与直接客户的关切与利益**，而对**间接利益相关者（indirect stakeholders）**的影响关注不足 [115]。

此外，这种“短视性”的范围设定也可能导致对某些特定类型风险的忽略，尤其是那些只有在人工智能系统与**长期存在的权力不对称结构**发生交互时才会显现的风险。同时，这种评估模式也往往**不重视将审计结果传达给公众**，也**缺乏推动非技术性干预或集体行动**的积极意愿 [16]。

3.1.2 速度（Speed）

在算法审计（algorithmic auditing）领域的早期研究中，审计过程曾被积极评价为“必然是枯燥、缓慢、细致且系统的”工作 [114]。然而，放到今天的现实语境下，这类表述似乎已显得不切实际。

随着各类组织试图将**法律要求与监管规范**转化为可执行的操作流程 [86]，评估、审计与审核工作的节奏被期望能够与**人工智能系统的高速迭代开发保持同步**。这一对“速度”的期待，实际上**抑制了对那些耗时较长的方法的采用——**尤其是那些需要深入分析技术所处的社会、经济与政治语境的方法与路径。

简言之，为了追求“快速评估”，许多本可以揭示深层风险的分析方法被边缘化，使得评估工作趋向表层化与技术中心化。

3.1.3 规模（Scale）

人工智能评估与审计方法的**标准化趋势**进一步强化了一种假设：即算法系统应当**在其自身的技术框架内进行评估——**换言之，主要通过**技术性、统计性、数据驱动且可扩展的方法**来进行分析 [117]。

随着社会各界对“清晰定义”与“可标准化”的审计要求日益增多 [86]，这种趋势持续加剧。评估实践逐渐围绕诸如基准测试（benchmarking）、偏差检测（bias testing）以及各类统计分析工具展开，而这些工具往往**在市场中被买卖、流通和商品化**。

事实上，Costanza-Chock 等人 [31] 的研究发现：在其所访谈的审计从业者中，大多数更倾向于使用**量化方法**而非**质性方法**；同时，**不到一半**的受访者将**利益相关者的参与**视为人工智能审计的重要组成部分。

更进一步，Raji 等人 [112] 指出，为追求**可扩展性**与**标准化**，这些评估技术常常陷入**过度泛化**，从而导致**建构效度（construct validity）问题——**即评估指标与方法未能真实反映其声称要衡量的风险或影响。

3.1.4 场所（Site）

由于人工智能审计产业的主要参与者大多集中在**全球北方（Global North）**，评估与审计工作的执行地点也通常围绕**该地区的终端用户与利益相关者**展开。这种地理集中导致非洲、拉丁美洲（含南美与中美）以及中亚等地区在人工智能伦理与治理领域中**代表性严重不足** [69]。

然而，**全球南方（Global South）**并非是数据驱动技术发展过程中的“沉默接受者”或“被动对象” [125]，相反，它在全球人工智能/机器学习（AI/ML）研究与产业生态中扮演着关键角色，包括：

南方国家本土的 AI/ML 研究与创新实践；

大规模商业化数据提取的主要来源地；

数据标注与注释的劳动市场；

人工智能系统的测试试验场所（beta-testing sites）；

稀有矿产与原材料的主要供应地 [106]。

因此，这些地区**本应是评估工作的重要现场**，却因结构性忽视而被“隐形化”。

总结：当前人工智能评估与审计的整体生态正面临一个日益趋窄的趋势，不论是在**范围、速度、规模**还是**评估场域**层面。这种收窄倾向正将原本应是一个**充满研究活力与争议空间的实践领域**，转变为一组**标准化交付成果** [141]，其政治性也因此被消解，更易受到**企业利益的捕获** [53]。

在下一个章节中，本文将探讨**将历史分析引入评估与审计流程**的潜在价值，并指出历史方法**如何有助于对抗上述收窄趋势**，从而回应人工智能评估作为**公众利益监督机制**的初衷。

4 将历史作为人工智能评估的方法

与人文学科和社会科学中的其他分析方法类似，**历史分析**的目标在于深化我们对人工智能系统的**设计、开发、整合、（再）使用与流通过程中的社会、文化、经济与政治因素**的理解。然而，历史分析又作为一种**独特的社会技术分析路径**而有所不同，原因在于：

它将“过去”视为**批判性理解当下的知识场域** [143]；

它特别关注**档案记录**以及**当代人工智能系统的深层历史根源**。

正因其具有独特的**时间取向**，历史方法的引入能够**拓宽传统人工智能评估的视野**，促使我们回顾多个历史时刻、物件与实践，将它们视为构成当代 AI 系统的历史条件 [107]。

进行历史分析时，可参考多种来源材料，包括：

档案记录

口述历史

历史文物

已有的历史学术文献

这些材料通常保存在**图书馆、博物馆、社区中心、机构档案、私人收藏**及各类历史遗址中——这些空间不仅承担着信息储存的功能，也通过代际传递构成了历史知识的重要来源，从而为历史分析提供了现实起点。

在具体研究路径上，历史分析可采取不同取向：

地方史（local history）或微观史（microhistory）方法，聚焦某一特定社区、地点或历史时期；

也可采取**长时段研究（longue-durée）**方法，关注那些在长时间尺度上造成深远影响的结构性变迁与历史规律。

哪种来源、场所与方法最适合具体评估任务，将取决于：

评估者的实际可用资源与获取历史材料的能力；

所评估的**技术类型**与相关**利益相关方**；

以及评估者自身的**知识立场与方法论位置**（positionality）。

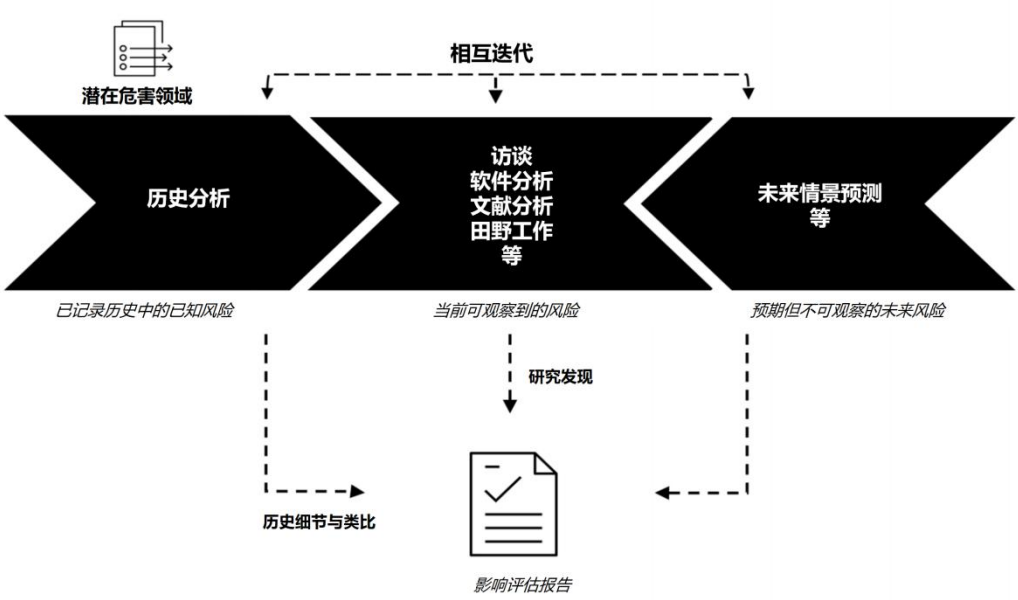

图 1：**融合历史分析的评估框架高级概览图**。
该图展示了不同评估方法在时间维度上的关注焦点、它们在迭代过程中的互动方式，以及历史分析所预期产出的类型。

4.1 历史分析在评估协议中的适配性

历史分析本身并**不是一种全面的评估方法**。为了能够完整评估并预测一个人工智能系统的风险与社会影响，评估者必须具备对当下实践、话语、运行机制、技术能力、限制条件与已观察到的影响等多个方面的深入理解。因此，历史分析最有效的方式是**与其他社会技术评估方法协同运作**，包括但不限于：访谈、民族志田野调查、用户研究、模型测试、数据分析、以及参与式干预等。

图 1 展示了一个我们将历史分析嵌入到社会技术评估过程中的实例（详见第 5 节）。以下我们将概述历史分析如何以三种广义方式，嵌入多种社会技术评估框架中，从而**增强其在范围设定、风险发现、研究设计、评估实施与结果传播各阶段的能力**。

我们具体建议将历史分析纳入评估协议中，以用于：

扩展**危害识别阶段（harms discovery phase）**的视野；

指导研究问题的设计、测试流程和干预实践；

支持**发现的实质化表达与结果传播**。

---

1. 扩展“危害识别”阶段

将历史分析引入评估流程最有效的时机是在“危害识别”阶段 [16]——即评估对象、可能受影响人群以及潜在危害类型被初步界定的阶段。

原因在于，历史资料常常包含关于**过往技术相关危害事件**的信息。这些信息可以与现有的风险分类体系结合，从而识别出 AI 系统或其部署场景中需要被纳入评估的新风险维度或潜在脆弱点。

---

2. 指导研究问题、测试设计与参与式过程

历史分析还可以用于**塑造评估过程中所提出的研究问题**。例如：

如果历史记录揭示了某些根深蒂固的假设、文化规范或刻板印象，那么就可以据此优化访谈问题，用以探查这些规范是否仍然存在于 AI 系统的设计者与用户之中；

红队测试（red teaming）与对抗性测试也可基于历史先例设计，以检验模型是否重演类似历史问题；

在参与式研究环节中，历史洞见还能**作为与受影响社群交流的关键资源**。Katell 等人 [74] 在其参与式研究中发现，被 AI 系统影响的社群对类似技术在过去造成的社会影响很感兴趣，这有助于他们更清晰地表达当下的具体诉求与担忧。

---

3. 加强发现的表达力与传播力

历史分析的第三种用途是作为一种**阐释与传播评估发现的有效工具**，便于与不同听众与平台沟通 AI 审计的意义与严重性。

历史案例研究可以帮助说明为何某些AI系统的局限性比其他问题更具社会风险，从而有助于缓解措施的优先级排序；

历史类比与对照也能为某些“尚未发生的”或“长时延迟型”的潜在后果提供合理预判与干预依据 [27, 33, 81, 160]。

在接下来的章节中，我们将通过一个案例展示我们是如何在实践中**以这三种方式**运用历史分析来界定并执行一次人工智能系统的评估。我们希望，这个实例能够说明：历史分析如何**可操作地整合进评估协议中**，并为那些希望将“过去的洞见”嵌入到“当下与未来 AI 系统研究”中的评估者、审计者与评价者，提供一个可行的初始框架。

第 5 节 案例研究：虚拟代理系统的历史分析

本节案例来自我们此前对某大型科技公司**虚拟代理系统（virtual agent system）所进行的一项深入技术评估。鉴于技术的保密性，我们不会披露该评估的完整结论，而是聚焦于历史分析是如何作为方法论嵌入到整个评估流程中的**。

通过这个例子，我们希望展示一种可能的路径：如何系统性地开展历史调查，并在评估框架中有效地运用其产出。这个案例体现了历史分析**在实际评估任务中具备可操作性与策略价值**，为其他评估者提供了一种实践范式。

5.1 背景

5.1.1 机构背景

本案例研究来自一项**社会技术评估**工作，旨在评估某虚拟代理系统的使用方式及其潜在影响。在本文中，**虚拟代理**指的是**具有人类外观特征并可与终端用户进行实时交互的 AI 驱动对话代理系统**。

本次评估的目标包括：记录该系统已知或可能存在的负面影响风险（尤其是**偏见问题**），并就如何在部署前优化技术设计、加强组织内伦理机制的整合提出建议。需要说明的是，这项评估并不打算替代由独立监管机构进行的外部审计，而是如 Knowles 和 Richards [78] 所提出的那样，被视为一种“**从超越部件层面思考 AI 影响的良好起点**”。

评估团队由若干全职研究人员和一名暑期研究生实习生组成，均隶属于一家大型科技公司。团队的专业背景涵盖**技术、社会科学与人文学科**，在评估开始前，对于虚拟代理系统的经验和知识水平不尽相同。

在为期 **16 周** 的评估周期内，研究人员开展了一个包含 **六个阶段** 的评估流程（见图 1），并采用了**多方法、多学科交叉的研究策略** [11]。各阶段以交错推进的方式展开，便于及时共享阶段性成果，从而对后续研究的范围与重点进行调整。

出于这一需要，**历史分析从第一阶段便开始介入**，目的是将待评估的虚拟代理系统置于其**更广阔的历史脉络**之中，揭示其所继承的社会、政治与经济维度上的技术背景。

每周，历史分析的**初步发现会向整个评估团队进行汇报**，以呈现当前阶段的研究成果、发现新的关注领域，并根据新获取的信息对评估协议做出必要调整。这一工作机制促成了历史分析与整个六阶段多方法评估流程之间的**持续迭代与双向反馈**。

5.1.2 方法与实施

本次历史分析采用了**叙述性文献综述（narrative review）**的方法 [50]，回顾了与虚拟代理相关技术的已有历史文献，包括人工智能（AI）、聊天机器人、机器人与自动机，以及某些具体的技术、方法或能力，如面部识别、情绪识别和情感计算（affective computing）。

开展叙述性综述的动机在于围绕以下三个主题进行主题分析：

种族、族裔、性别、能力、国籍、阶级等因素在相关系统中是如何被体现的；

情绪的不同概念与理解方式在心理测量和情绪识别技术中是如何出现的；

劳动的观念在对话代理与自动化客服系统的营销、设计和实施中是如何呈现的。

之所以仅进行文献综述，主要是出于时间限制，特别是希望尽快完成评估中的危害发现阶段。

本次文献综述聚焦于 **15 本科学史与科技研究领域的重要期刊**中的**同行评审文章**。我们使用表 1 中列出的关键词，在文章标题和摘要中进行检索，并由作者分别独立筛选文献。

被纳入的文章需同时满足两个条件：一是提供与虚拟代理相关技术的历史叙述，二是涉及如下主题之一：表征、身份与主体性；情绪、情感与软生物特征；劳动、资本主义与自动化。不涉及虚拟代理或与其无直接关联的 AI、自动化或数字技术（如工业制造领域）的历史研究文章未被纳入。

最终，共选取了 37 篇文章，用于探讨上述三个主题。随后又通过“滚雪球”方法从参考文献中识别出其他相关书籍和文章。每个主题下最终纳入的文献详见表 2，更多细节见附录 A。

Table 1: Historical Literature Review Keywords

1st term		2nd term
artificial intelligence	AND	representation
virtual agent		identity
chatbot		subjectivity
robot		race
automata		ethnicity
		gender
		ability
		class
		emotion
		affect
		labor
		capitalism
		political economy

Table 2: Historical Literature Review

Theme	Relevant Articles and Books
Representation, Identity and Subjectivity	[119], [144], [44], [84], [139], [41], [32], [66], [129], [56], [23], [140], [57], [97], [124], [72], [105], and [136].
Affect, Emotion, and Soft Biometrics	[20], [92], [144], [1], [142], [122], [156], [131], [132], [35], [155], and [63].
Capitalism, Labor, and Automation	[87], [6], [105], [148], [149], [70], [41], [84], [9], [39], [37], and [8].

5.2 历史发现

通过我们的分析，我们识别出了一些**影响虚拟代理相关技术的设计、营销与实施方式**的共通性实践、叙事框架、基本假设与权力结构。这些要素具有深刻的历史根源，并在可能的范围内被放置在其相应的**社会、文化、经济与政治语境**中加以理解。

下面，我们总结了本次历史分析中获得的**关键洞见**，并说明这些洞见如何影响了我们对虚拟代理系统中**潜在危害领域**的理解。同时，我们也展示了这些历史发现如何被用于构建后续研究流程中的问题设定，并最终在评估报告中为部分结论提供支撑。

5.2.1 表征、身份与主体性（Representation, Identity, and Subjectivity）

现有历史文献表明，自 18 世纪欧洲**首批自运行机器（automata）普及以来，从自动机、机器人到如今的 AI 代理系统，这些技术长期以来在其设计、审美呈现与拟人化方式**中，持续嵌入了特定的**种族、性别、阶级和残障方面的刻板印象与社会规范**。

例如，Crawford 对 18 世纪早期自动机的研究 [32]，尤其是对 David Roentgen 所制造的**大键琴和洋琴演奏自动机**的历史分析，指出这些自动机反映了当时**欧洲资产阶级文化中有关性别的观念**，并作为“性别意味”的具象物，成为**女性气质如何被定义与争议的载体**。

类似地，Frumer 关于 1920 年代大阪制造的机器人「学天则（Gakutensoku）」的研究 [41] 也揭示，该机器人设计上融合了“多种族的外貌特征”，体现了日本当时的**帝国多元文化主义（imperial multiculturalism）理念。这表明文化与政治愿景在自动化技术的造型、美学与拟人化设计**中扮演着关键角色。

而对更近期技术的研究同样指出，**长期存在的文化规范与刻板印象**依旧深刻影响着当下的 AI 系统。Lingel 与 Crawford 关于 AI 助理与“秘书”历史之间关系的研究 [84]，以及 Phan 关于 AI 助理与家政服务历史之间的研究 [105]，均揭示了这一点。

Lingel 与 Crawford 指出，**秘书这一角色作为“能干、支持性强、随叫随到、女性化的下属”**的形象，在早期 AI 助理的“女性化、白人化、受过教育的语调与形象”中得以复刻与延续 [84]。

由此我们可以看到，从自动机到机器人再到 AI 代理，这些系统在有意、无意，甚至不可避免的情况下，都在表达特定的**性别、阶级与种族身份** [66, 105, 129]。这类表征往往体现了社会主流对“人类差异”与“人类行为”的理解，从而在现实中**复制并加固了排斥性与歧视性的刻板印象与社会规范** [44]。

5.2.3 资本主义、劳动与自动化（Capitalism, Labor, and Automation）

科技史与科学史学者早已指出：**自动化技术与人类劳动的表征之间存在长期的历史纠缠**。例如，Jones-Imhotep 对自动机历史的研究 [70] 回溯了 18 世纪使自动化表演看似可信的社会条件与修辞手段。他发现，这些技术系统中“自主性”与“能动性”的幻象，其实依赖于对**使这些系统成为可能的劳动的抹除——**尤其是对奴隶制历史所构成的劳动观的依赖。在这种文化语境中，18 世纪的观众被“训练”成能够**有意识地忽视**那些明明存在于眼前的劳动者。

类似地，Atanasoski 与 Vora [7] 研究了“后劳动社会想象”中所谓的**替代性劳动（surrogate labor）**，尤其是那些宣称**自动化能够“解放人类”脱离重复性、无创造性的工作**的技术愿景。他们指出，这种“解放”承诺常常是建立在**对某些形式的劳动进行去人化（dehumanization）之上的，而这些劳动往往具有种族化、性别化、以及社会贬值的特征**。

到了 20 和 21 世纪，这些结构性遗产在当代 AI 系统中依然可见 [4, 7, 48, 105]。例如，Phan Phan [105] 指出，像 Amazon Echo 这样的 AI 助理系统，“持续地被建构为理想化的家庭服务形象，**模拟着 19 至 20 世纪美国家庭中主仆关系的结构**”。再如 Alfred 与 Amazon Mechanical Turk 这样的技术平台，也以“替代效应（surrogate effect）”为消费者制造出一种**无需面对劳工反抗即可享受技术自由的幻想** [6]。

这种对自动化“革命性潜能”的想象持续地影响着人们对**哪些劳动可以被视为“过时”、从而应当被取代**的判断，并进一步助长了一种**不切实际的期望**。这些期望可以追溯到 **16、17 世纪殖民时代对劳工的规训技术**，以及 **18 世纪伦敦工厂中工业机械的乌托邦式“后劳动社会”承诺**。

5.3 与整体评估的关联性

我们的历史分析所获得的洞见被用来启发并指导后续社会技术评估方法中的研究问题制定（包括由评估团队其他成员进行的访谈、软件测试、启发式评估、数据分析以及未来工作坊等）[11]。

例如，自动机、机器人、聊天机器人和对话式 AI 系统的历史中反复出现的**种族、性别、阶级与能力的刻板印象及文化规范**，促使评估团队意识到：必须深入探讨虚拟代理系统在**具体设计、美学化与人格拟人化**等方面所体现的社会文化维度。为此，团队对设计者与终端用户界面进行了启发式评估，并在基于现有虚拟代理设计选项开发的探针（probe）基础上组织了未来工作坊，重点考察了两个方面：（a）虚拟代理的特征、技能与个性设计中所涵盖的选项范围；（b）是否存在关于种族、性别、族裔、能力、阶级等方面的刻板化或贬损性表现。尽管许多 AI 系统呈现出**“白人、女性、中产阶级”的美学倾向**早已为人所熟知 [24]，但历史分析中的某些视角——如日本机器人“学天则”（Gakutensoku）受帝国多民族理想影响而呈现“混合族裔”外观的案例 [41]——进一步促使团队从批判视角审视更多样化与混合型的设计选择。

类似地，关于情绪与情感的科学理论历史，尤其是 **19 世纪心理测量学对面部情绪识别（FER）算法的影响** [131]，以及 **20 世纪中后期对“基本情绪”概念的普遍化处理** [156]，也引导团队聚焦于虚拟代理系统中的情绪识别与情感计算功能。团队在用户测试与系统训练数据的分析过程中，特别关注以下几个评估点：（a）情绪识别能力的准确性；（b）系统对不同群体情绪表达与理解的“规范性假设”；（c）是否存在关于某些社会情境下“恰当情绪表现”的先入之见。历史研究所揭示的——人们基于外貌与社会背景而形成的情绪判断中的规范性假设——促使团队将评估拓展到技术功能之外，更关注**虚拟代理在广泛社会文化条件下的实际表现**。

此外，关于自动化劳动背后的**文化、社会、经济与政治力量**所塑造的文化想象，也为团队提供了考察虚拟代理系统与劳动问题的背景视野。16 世纪至 18 世纪间，随着**殖民地劳动规训技术的兴起**和**工业机械推动的“后劳动社会”乌托邦幻想**的扩散，出现了“替代性劳动（surrogate labor）”的文化想象 [6]。这些历史背景促使团队在分析宣传材料、访谈从业者、以及开展未来工作坊以评估虚拟代理在工作场所部署反应时，聚焦以下两个方面：（a）设计理念与话语体系中是否包含对“理想工人”的特定构想；（b）虚拟代理自动化任务后对劳动体系可能产生的影响。在这一点上，历史分析推动团队去挖掘话语中潜藏的、历史积淀下来的**关于情绪管理、举止规范与职业化标准的期望**。

除了如上所述用以识别评估议题，历史分析还在**最终评估报告的撰写与成果传播**中发挥了作用。具体而言，团队借助历史细节、背景、类比与对照来：（a）解释为何某些风险轴线成为评估重点；（b）论证为何部分风险被评为更值得担忧；（c）借助历史类比说明技术能力发现的潜在社会风险；以及（d）为特定缓解策略的优先级排序提供论据。

6 讨论

“‘过去’从未真正过去。”——引自 [145]

我们的工作让我们收获了关于**将历史分析作为一种方法融入 AI 评估与审计实践**中的诸多经验教训。这些经验来源于我们对某虚拟代理系统进行的为期 16 周的内部评估，以及我们对该经历与第 4 节所设定目标之间的对照与反思。

具体而言，这些目标包括：

更深入理解虚拟代理系统的历史生成背景，重点记录相关技术与科学实践已知的社会风险、影响与含义；

利用这些已记录的风险来指导后续评估阶段的范围设定、进一步调查的关注重点识别，以及研究问题的提出；

协助呈现评估工作中其他六个工作流所产生的发现的社会影响与现实利害关系，用于最终报告撰写与沟通。

在接下来的内容中，我们将围绕上述目标，总结我们在实践中遇到的**具体益处与挑战**，并将这些经验放置在当前更广泛的 **AI 审计行业所面临的激励结构与制度限制**的背景下加以分析。

6.1 历史分析的益处

6.1.1 扩展技术生命周期之外的时间尺度

当前有关人工智能评估的框架、方法、工具和标准的研究，普遍强调应覆盖整个技术系统的生命周期——即从技术构想到其重用或退役的全过程。这一呼吁纠正了以往评估仅聚焦某一时间节点的局限性，具有重要意义。然而，我们在此次评估中引入历史分析的实践经验进一步推动我们认识到，仅仅围绕技术生命周期本身进行分析仍然不够。

将评估的时间维度扩展到技术生命周期之外，意味着必须认识到：几乎所有人工智能系统实际上都是在既有的技术、基础设施、软件和硬件之上构建起来的，而这些技术元素本身也各自拥有独立的历史。这种历史延续性不仅体现在物质层面，更体现于前代技术和科学范式所留下的结构性与组织性印记，这些印记蕴含着特定的社会、文化与政治动因、假设与影响[36][18][119]。

以虚拟代理系统为例，为了还原其完整的历史脉络，我们必须围绕多个不同的关联技术与科学理论开展历史调研。这些技术和理论并非该虚拟代理系统生命周期的直接组成部分，却构成了其深层的技术与观念渊源。我们回溯了从 18 世纪自动机到当代 AI 助手、从 19 世纪心理测量学到当代面部情绪识别技术（FER）、

从 20 世纪关于工人起义的焦虑到“后劳动社会”的美好承诺等多个轨迹。这种广谱式历史调查促使我们：

探讨虚拟代理的美学设计如何延续并反映了有关种族、性别、阶级和能力的文化规范；

质疑自动化情绪识别的“准确性”“普适性”与“客观性”叙事；

拓展对“替代工人”与“理想劳动者”想象的批判性反思。

历史分析所带来的这一“长时间尺度”视角，极大拓展了我们对潜在危害的发现范围，也提升了整个评估工作的深度与广度。

6.1.2 将风险识别与干预策略推进至“实验室之外”

我们在历史分析中获得的诸多见解，集中体现了虚拟代理系统在文化层面的影响。在评估工作中，这促使我们必须超越仅关注实验室内部实践的狭隘视角，转而将研究和干预指向更广阔的社会语境与权力关系[43]。

在本案例中，“走出实验室”意味着将评估工作拓展到那些影响虚拟代理系统设计、推广与使用的社会情境与权力结构之中，而不仅仅聚焦其技术组件。例如，在我们回顾自动化与劳动相关技术所引发的危害时，发现“替代性劳动”（surrogate labor）这一文化想象是一种极具影响力但又难以捕捉的中介机制。它在历史上深刻地塑造了自动化技术的劳动影响，但往往并不直接体现在数据集或算法系统中，而是隐含于产品演示、客户交流、市场营销材料、以及讨论论坛等非技术环节。

正因如此，历史分析的引入帮助我们拓展了传统评估中使用的“输入”与“输出”范围，强调了一种更具关系性与文化敏感性的人工智能评估方法[116]。这使我们能够在更完整的社会、文化、政治与历史脉络中审视潜在风险，而非仅仅依据某种固定的量化指标[73]。

此外，我们对“自动化文化”延续性的关注也让我们在评估过程中兼顾了技术的长期影响与短期表现。在此视角下，我们的发现能够更敏锐地意识到各类风险与影响是如何与殖民主义、资本主义剥削等历史结构交织在一起的——而这也意味着其后果并非平均分布，而是沿着地区、社会与群体间的不平等裂痕呈现出差异化的体现[106]。

6.1.3 将前瞻性治理扎根于历史语境之中

历史分析的另一个关键价值，在于其能够为其他评估方法与干预策略提供背景支撑。当人工智能评估与审计旨在预判未来潜在风险时，历史方法所提供的脉络信息，能够帮助构建更为现实、可预见的近未来影响图景。正如计算史学者 Michael Mahoney 所言：“历史通过厘清我们此刻所处的位置以及到达这里的路径，帮助

我们理解可能的前行方向”[91]。而据 Coopersmith 和 Daemmrich 的观点，历史还“能为我们提供洞见，揭示何种条件能够促成某些政策的成功、而另一些则注定失败”[30]。

在本次针对虚拟代理系统的评估中，我们所开展的历史调研生成了若干“历史片段”，这些片段后来被用于启发用户设想潜在未来场景，或作为用户响应近未来乃至远未来虚构情境的出发点[75]。例如，一则关于“友好”或“易于共鸣”的机器人在日本帝国政治宣传中被使用的历史故事，为我们设计的虚构未来情境注入了真实质感——该情境设想了未来虚拟代理系统如何被用于企业多样性宣传。与其凭空臆造未来场景，不如通过历史分析为前瞻性探索提供现实语境的支撑。

简言之，历史分析使我们得以将对未来潜在危害的预判工作，锚定于真实历史语境之中，从而提升前瞻性治理的可操作性与可信度。

6.1.4 系统特征的再政治化

历史分析为整个评估流程提供了一个关键契机，使研究问题得以超越人工智能系统的技术能力本身。在当前强调统计分析[117]、审计标准化[86]以及量化方法优于质性方法[31]的趋势背景下，聚焦历史视角使我们能够记录介入人工智能系统设计、开发与使用过程中的社会、政治与经济背景。这一聚焦路径有助于将系统的特征、能力与局限“重新政治化”，即将个别设计选择与长期存在的文化规范与结构性条件联系起来进行考察。

例如，我们对科学情感理论的回顾表明，面部情绪识别算法中某些表现偏差或技术局限，实则可以追溯到心理测量学的历史，它们并非是技术尚不成熟所造成的“中性”限制，而是带有历史负载的偏见产物。再如，通过历史分析，我们得以将虚拟代理系统在外观、审美化与拟人化方面的设计选择重新政治化，指出其实质是在延续一种长期存在的、带有种族化、阶级化、性别化与健全人中心主义色彩的“专业主义”文化刻板印象，而非某些孤立的设计个体偶发为之的结果。

在上述两个例子中，历史方法都提供了批判的契机，使我们得以对抗将人工智能危害简化为纯粹技术问题或孤立设计选择所带来的偏狭视角。

6.2 考量与局限

6.2.1 时间与资源限制

某些类型的历史分析（例如基于档案研究或社区口述史的方法）往往耗时较长。在当前要求快速完成人工智能系统评估的背景下，这类历史分析常常被视为不切实际。在我们所进行的历史分析中，所处的企业环境也确实存在这样的时间压力。因此，我们选择主要依赖已有的历史文献来开展分析。

然而，这一选择也带来了额外的问题。例如，许多科技史与技术史期刊仍受限于付费墙，对于无法通过机构渠道获取期刊资源的评估人员来说，这可能构成实质性的障碍。我们在评估过程中也经常面临获取人文与社会科学期刊资源受限的困境。

因此，在倡导将历史方法更广泛地纳入人工智能评估与审计工作的同时，我们也希望能同步推动相关资源的公众可达性与开放获取。

6.2.2 专业能力与访问条件的考量

历史分析的开展通常需要具备一定的历史研究方法知识，这与其他定性方法、技术评估或合规性审核在人工智能评估中的应用情况类似。学术历史学者、档案馆员、博物馆策展人、社区历史工作者以及历史遗产保护从业者等，均具备开展 AI 系统历史分析的专业能力。在本研究中，我们依赖于自身接受过的历史学科训练来执行相关工作。

这一点为分析带来了积极的作用，但也可能由于受限于特定学术训练框架而在无形中限制了我们对某些风险与影响的识别范围。换言之，我们所“看见”的风险和影响，在某种程度上受到了历史研究方法自身视角和取径的影响。

除了专业能力之外，历史分析的有效性在某些情况下还依赖于对被评估技术的具体细节的可获取程度。虽然历史分析可以由内部或外部评估者执行，但当评估者能够接触到 AI 系统的组件构成、预期用途以及具体部署场所的信息时，分析通常会更为深入和严谨。

正如已有研究指出的那样，获取这类信息对于开展有效的人工智能审计与评估至关重要[22]。然而，在面对技术公司进行第三方外部研究时，信息获取往往具有相当的挑战性[118]。

6.2.3 历史分析的局限性

在开展历史分析时，必须意识到一个关键问题：关于技术的历史叙述本身就反映了叙述者的视角、政治愿景、文化假设以及社会想象[135, 145, 158]。正如 Trouillot 所指出的：“过去，或者更准确地说，‘过去性’是一种立场”[145]。换言之，事件本身与对事件的叙述之间往往并不完全等同，当我们依赖关于过去的叙述来评判当下的技术时，历史所承载的意义与影响就会变得更加敏感与关键。

因此，在应用历史分析时，需要格外谨慎，避免将某些相关的历史经验抹去，尤其是要注意将边缘群体与弱势社群的历史纳入分析之中。“从下而上的历史”、聚焦于普通人的历史、以及关于非西方/非欧洲地区科学技术发展的历史，都是实现这一目标的关键路径。

此外，还需重视历史类比本身所存在的局限性。过去与现在之间，尤其是不同地区与社区之间，往往存在重大的语境差异。在我们利用历史分析所得见解，并将其用于指导当下技术评估时，常常需要回到具体的情境细节中，以避免产生过度概括或错误类比的风险。

在这些情况下，将历史分析的发现与其他评估方法所得出的观察结果进行交叉比对，显得尤为重要。这样的反思性操作有助于保持分析的准确性与批判性。

6.2.4 评估工作的被俘获

尽管我们将历史分析方法引入 AI 审计与评估体系的努力，初衷是回应 AI 审计产业在方法论上日益趋于狭窄的趋势，但我们必须承认，这一工作本身仍是在一家科技公司的研究体系内开展的。因此，它无法脱离更宏观的结构背景来理解。

正如 Grill [53] 所指出的："尽管测试制度化的进程正在推进，它却往往被技术官僚所主导，而测试的本质其实是政治性的。" 我们的经验也确实反映出一个广泛的担忧：AI 评估实践已从其最初旨在实现公共问责的目标中偏离，这种偏离被认为削弱了"测试所承载的解放性承诺——也就是通过测试来揭示系统造成的那些大科技公司常常忽视的危害"[53]。

尽管我们尝试将本研究的价值定位在方法论层面的反思与启发上，但若要真正回归 AI 评估最初的公共利益愿景，实现其"解放性承诺"，恐怕仍需在大型科技公司的体系之外另辟蹊径。

7 结论

本文认为，尽管在对社会技术系统的批判性研究中，历史因素的重要性已广为认可，但将历史分析作为一种正式方法系统性地纳入 AI 评估与审计协议之中，目前仍属罕见。

将历史分析融入 AI 评估、审计与评价流程，并将相关发现分享给政策制定者、技术开发者、负责任技术从业者以及受影响的社区，有一个显著优势，即其**具有累积性和协作性潜力**。我们主张，历史分析可以通过以下几方面提供独特而有价值的视角与洞见：

首先，它能够**拓展评估的时间维度**，超越传统对技术"生命周期"的关注，提醒我们关注那些**并非立刻显现**、却深植于技术背后的**社会风险与影响**。

其次，我们展示了历史分析如何与其他社会技术评估方法协同使用，用以**提出研究问题**、为评估活动提供**背景脉络**，并**支撑最终发现的论证过程**。尤其值得注意的是，历史语境可以**凸显特定风险或影响的社会意义**，通过将其置于更广泛的社会、文化、政治与经济背景之中，使其变得更加可感、可理解。

总而言之，若将历史研究的视野、目标与价值置于 AI 评估的核心，将极大拓展可发现的危害领域，并能以更具批判性和深度的方式传达评估结果——从而有效抵制 AI 审计行业中日益狭隘化的评估方法，持续聚焦 AI 系统对社会的伤害及其**长期影响**。