【教育创新探索】

伦理、分析技术与注意义务

（加）斯蒂芬·道恩斯

（加拿大国家研究委员会，渥太华 K1A 0R6，加拿大）

肖俊洪 译

【摘　要】人工智能和学习分析引发了许多伦理问题，因此公平、正义和仁爱等话题重新引起了人们的关注。本文聚焦学习技术，对这些话题进行全面分析，调查人工智能和分析技术的应用情况，指出学界提出的相关伦理问题。文章接着分析在设计人工智能和分析技术的过程中需要从伦理角度做出决定的情况，研究相关行业伦理准则并概括影响这些准则的伦理理论，运用注意伦理哲学分析当下问题，文章最后围绕伦理实践展开讨论。

【关键词】分析技术；人工智能；伦理；注意；社区；文化

【中图分类号】G434　　**【文献标识码】**A　　**【文章编号】**2096—1510（2024）03—0001—18

导读：文章是慕课和联通主义的始创人之一斯蒂芬·道恩斯（Stephen Downes）历时三年的研究成果，2023年秋应译者之约整理成这篇文章。

近年，AI伦理成为一个研究热点，但是，很多饱受诟病的问题依旧存在。为了破解这种局面，我们往往把解决问题的希望寄托在制度、准则、规则、原则、标准、规范等“大道理”上，虽然未曾完全取得预期的效果，但是我们依然乐此不疲。道恩斯尝试“另辟蹊径”，他提出的AI伦理新观念可概括为：伦理问题是文化问题；AI伦理应该是一种和谐伦理。译者的理解是，和谐伦理是一种参与式文化，体现的是AI利益相关各方的自觉伦理意识和素养，这与“制度化”的“大道理”形成鲜明对比。

“分析技术”（analytics）这个术语涵盖的范围要比“学习分析”（Learning Analytics）大得多，后者仅局限于指分析技术的具体应用。“分析技术”和“AI”在本文中经常互换使用。

文章开宗明义阐述了作者与众不同的伦理观：伦理旨在给人带来快乐，然而我们却往往以为伦理关乎的是对与错，因而把主要精力放在如何防止做坏事和错事上，而不是如何体验伦理之乐。正因如此，传统的伦理建立在原则的基础上，即“我们‘应该’或‘应当’（或‘不应该’或‘不应当’）”做什么，而非发自内心的“我们做好事时那种温暖和成就感”，“没有反映我们的同情心、正义感和善良”。今天，AI也面临相同窘况，这正是激发作者开展这项研究的初衷。

文章扼要介绍了AI的六大用途，进一步指出“分析技术伦理特别复杂，因为不管它正常运作与否都会引发伦理问题”，同时从五个方面展开剖析。

文章指出既然“我们能够用技术制造仇恨，或许我们也能够用技术制造关怀”，因此“必须思考如何做出这些设计和使用方面的决定才合乎伦理。”作者从七个方面阐述“涉及分析技术和AI的所有不同过程以分析我们在设计和使用这些系统时所做出的决定”是如何可能产生伦理影响的。业（学）界认为要通过制定伦理准则规范AI伦理。伦理准则经常是从普适性或人权等角度制定伦理原则的，而作者认为要多从务实的角度考虑，平衡风险与回报。文章分析了从美德和品格、义务、后果论和社会契约等角度理解伦理所碰到的问题，认为应该告别传统的伦理观，因为传统的伦理原则难以解决当下这个纷繁复杂世界所面临的问题。

[DOI编码]10.19605/j.cnki.kfxxyj.2024.03.001

基于这种认识，文章从五个方面对注意义务（Duty of Care）展开阐述。作者认为注意义务作为一个法律概念具有去人性化和功能性的特点，从伦理的角度看“有一些重大局限”。相比之下，作为源于女性主义认识论和女性主义伦理学的一个伦理概念，注意强调的是“一种真诚交流和关注另一方的需求的关系，不是死板和程序性的关系。”建立在注意伦理（Ethics of Care）基础上的“注意教学法”（Pedagogy of Care）能够激发学生的创造力，而注意伦理也“会重塑分析技术在学习中的角色”。伦理知识是一种隐性知识，一种感觉。因此“我们是在感觉或体验明辨是非的过程中学习伦理的，而不是从完美的一般原则中学习伦理的”；注意伦理“更像是一种情操或情感，更接近于一种平衡感”“我们在某个场合的反应取决于我们的伦理背景，是多种因素同时作用的结果，而不是因为制定了那些博眼球的大原则的结果。”这个意义上的注意是解读伦理的另一条“蹊径”。

文章基于注意伦理从五个方面讨论了学习分析中的伦理实践。“AI谈不上‘应用’伦理，没有哪一个人或集体要为伦理结果‘承担责任’……认为我们能够采取干预措施以研制出‘合乎伦理的’或‘可以解释的’AI似乎有误导之嫌”，相反“应该明确我们希望AI往哪个方向发展、AI应该服务哪个方向”“才能‘驯服’这种神秘技术”——这是有悖于当下主流话语的观点。作者分析了监管和管治的不足之处，认为采用“伦理实践框架”优于监管和管治。但是，管治和伦理框架不一定得到严格遵守和执行。基于此，作者提出要培养伦理社区，“把AI当成社区的一个遵守伦理的成员”，因为归根结底，伦理问题是文化问题。文章还讨论了伦理与文化的关系，认为参与式文化“能够重塑伦理的重点”。最后作者认为AI伦理应该是一种和谐伦理（Ethics of Harmony），是“一种注意、倾听和理解其他人缺乏什么、需要什么和能够提供什么的”、旨在促进公共利益的开放性伦理。

文章最后模仿丹麦哲学家索伦·奥贝·克尔凯郭尔（Søren Aabye Kierkegaard）的《非科学的最后附言》（Concluding Unscientific Postscript）一书书名，用“非伦理的最后附言”作为最后一节的标题，其寓意是：我们可以通过本质上不属于伦理学范畴（如文化与和谐）的过程实现伦理AI这个目标。

本文不乏一些另类，甚至可以说是较为激进的观点，这是道恩斯一贯的学术风格。希望这些观点能为创新AI伦理研究的方法、视角和思路提供有益的启示。

衷心感谢作者对本刊的信任和支持！

一、伦理给人快乐

伦理（Ethics）应该给我们带来快乐而非恐惧，其关乎的不是何为错而是何为对。伦理让我们有可能过上最美好的生活并怀有高尚、美好的抱负；伦理是我们实现这个目标的途径和工具。我们投入太多精力防止做坏事和错事，而这些精力本该用于努力创造美好和正确的东西。

同样的，应用学习分析（Learning Analytics）的最好结果不是防止发生伤害，而是创造美好。技术能够代表我们最好的一面，包括希望、梦想和抱负，这是技术存在的价值。然而，“古典技术哲学家给技术在当代文化中的角色描绘了一幅极其悲观的前景”（Verbeek, 2005, p.4）。我们赋予技术什么内涵？我们期望技术发挥什么作用？这些问题在分析技术（Analytics）领域显得尤为重要。

乍一看，伦理似乎关乎“对”与“错”（Mackie, 1983; Pojman, 1990），或如同亚里士多德所言，可能与美德和品格相关。不管是哪一种观点，伦理通常被认为是指我们“应该”或“应当”（或“不应该”或“不应当”）采取什么行动。

但是，笔者认为伦理是建立在感知（perception）而不是原则（principle）的基础上，源自我们做好事时那种温暖和成就感，反映我们的同情心、正义感和善良，因此，是发自内心的，不是能言巧辩或严厉训斥的结果。我们不遗余力地编造各种措辞和原则，好像能够以此说服他人遵循伦理。但是，有伦理的人不需要这些东西，而没有伦理的人则不会受到它们的影响。

分析技术的情况也一样。今天的人工智能（Artificial Intelligence, AI）引擎不是建立在认知规则或原则的基础上，而是通过大量情景相关的数据进行训练的，因此，它们无法做出伦理判断，而且我们也难以用简单语言告诉AI应当或不应当做什么。如同传统技术哲学一样，分析技术伦理的文献也表达了对人际关系疏远和向技术“俯首称臣”的担心。因此，我们既看不到分析技术好的一面，也找不到防止其危害的最好方法。

分析技术是一个崭新的领域，只有几十年的历史。然而，它所应对的是数百年来困扰着哲学家的

那些问题。当我们问何为对、何为错时，我们也是在问如何辨别对与错，如何领会对与错的区别并应用于日常生活中。人如此，分析技术也如此。

二、AI的用途

本文聚焦分析技术在学习和教育中的应用（通常被称为“学习分析”）。学习分析旨在提高学生成功的机会（Gašević, Dawson, & Siemens, 2015）。于是，在成立学习分析学会（Society for Learning Analytics）时西蒙斯把它定义为“测量、收集、分析和报告有关学习者及其背景的数据以了解和优化学习以及学习发生的环境”（Siemens, 2012）。

从分析技术伦理的角度看，采用一个宽泛的“学习分析”定义可能是最明智之举。例如：联合信息系统委员会（Jisc）《学习分析实践守则》（Code of Practice for Learning Analytics）中的定义（Griffiths et al., 2016）是使用学生以及学生活动数据“帮助学校了解和改进教育过程，向学生提供更好的服务”（Sclater & Bailey, 2015-2023）。本文中的“AI”或“AI和分析技术”采用的是这个定义。

虽然AI随着2022年秋生成式大型语言模型的发布而骤然走红，但它并不是新生事物，可以追溯到图灵（Turing, 1936）提出的概念、纽厄尔等（Newell, Shaw, & Simon, 1959）的“通用问题求解器”（General Problem Solver）和罗森布拉特（Rosenblatt, 1958）的“感知机”（Perceptron）。当然，所有这些都未曾达到通用人工智能的水平（General Artificial Intelligence）。AI这个术语指的是它所涵盖的诸多技术的宏大目标而非实际结果。因此，我们把大型语言模型称为通用人工智能时，并不是说它们是通用人工智能，而是说它们是实现这个目标的更大规模研究计划的一部分。

这个AI研究计划发展势头越来越强劲，其算法和模型已被融合到许多工具和过程中。 AI的用途大致可以分为以下六类，其中四类可见于相关文献中（Boyer & Bonnin, 2019; Brodsky et al., 2015），最后两类是我们增加的。

（一）描述性分析技术（Descriptive Analytics）

旨在回答“发生了什么”。重点是描述、发现和报告，包括各种来源数据的收集、过滤和整合机制，并生成各种可视化图表。这种技术可用于界定关键指标、发现数据需求、规范数据管理实践、提供用于分析的数据和向用户呈现数据（Vesset, 2018）。

（二）诊断性分析技术（Diagnostic Analytics）

旨在回答“为什么会发生这件事？”。对数据进行更加深入分析以发现模式和趋势，可以用于根据样本或训练数据中所发现的模式对某一数据进行推断，比如辨识、分类或归纳范畴、识别AI生成的脸谱（Li & Lyu, 2019）、情绪分析（Rienties & Jones, 2019）和自动评分（Lu, 2019）。

（三）预测性分析技术（Predictive Analytics）

旨在回答“将会发生什么事情？”。利用数据预测未来事件，可用于支持资源规划和事件响应（Drew, 2016）、学习设计（Rientes & Jones, 2019）和学习咨询指导（O’Brien, 2020）。当代分析技术的预测算法是基于考虑各种环境数据的模型，因此预测可能是数以千计变量相互影响的结果。

（四）规定性分析技术（Prescriptive Analytics）

旨在回答“我们如何能够使之发生？”。用于推荐行动措施，常见的是推荐学习起点内容或作为基于学习分析的学习路径一部分的内容。如支持学生个性化学习（Sclater, Peasgood, & Mullan, 2016）、自适应组建学习小组（Zawacki-Richter et al., 2019）、人员招聘（Metz, 2020）和决策（Parkes, 2019）。

（五）生成性分析技术（Generative Analytics）

旨在利用数据创造新东西。从某种意义上讲，它很像预测性和规定性分析技术，因为其推断不囿于提供给它的数据范围。但是，预测性和规定性分析技术需要我们行使能动性，根据分析结果采取行动，而生成性分析技术则是自主采取行动。随着诸如Stable Diffusion和ChatGPT的发布，生成性分析技术成为公众关注的热点。

（六）道义性分析技术（Deontic Analytics）

旨在回答“什么事情应该发生？”。这种问题越来越被寄望于分析技术，这是无法回避的。最近与之相关的一个问题是无人驾驶汽车，无人驾驶汽车也将会面临与菲利帕·富特的电车难题（Trolley Problem）（Foot, 1967）相同的问题。于是，微软研究人员研发了“定义问题测试”（Defining Issues

Test）系统，用以评估诸如GPT-3和ChatGPT这些新AI系统的伦理判断能力（Tanmay et al., 2023）。这些结果是程序员提前确定的，如专供优势家庭使用的汽车会把保护乘客而非路人预设为其伦理优先选项（Morris, 2016）。但是，并非全部伦理结果都会被预先编进程序。AI的伦理取向常常可能是其他优先考虑事情和活动的一个副产品。AI的本质和存在会驱动重大的社会变革。“AI技术在全社会引发了事关权力和控制的根本问题，预期随着社会转变成一个‘数字生活世界’，AI将给几乎每一个领域的人类活动都带来挑战。AI的用途与所引发的担忧涉及多个领域，包括从法律判决和维持治安到医疗保健、交通运输和军事活动等领域”（Liu et al., 2020, p.2）。

三、与AI相关的伦理问题

任何工具都不可能没有伦理风险。但是，分析技术伦理特别复杂，因为不管它正常运作与否都会引发伦理问题。纳拉杨把这些问题分为三类：分析技术有效运行引发的问题、分析技术尚不可靠引发的问题和分析技术应用似乎从根本上讲是错误的而引发的问题（Narayan, 2019）。我们另外增加两类问题：利用分析技术系统作恶以及与分析技术和AI使用相关的社会和文化问题。本节将呈现从文献和大众传媒中收集的问题。

（一）分析技术有效运行

“机器学习技术能在训练数据中发现模式，而不是依靠人工编程的显式规则，因此现代AI（几乎总是）能有效运行”（Liberman, 2019）。如上所述，分析技术用途广泛，包括简单识别、较深层次的诊断、预测、生成新内容和决策。很多情况下虽然我们不知道它是如何做到的，这不无好处，犹如魔术，观众不懂魔术原理更能享受其乐趣。但是，分析技术的精准性反而可能引发伦理问题。

监控：可能影响人权和自由（Shaw, 2017）。

追踪：“在公共场合被看到是一回事，被跟踪是另一回事”（Cavoukian, 2013）。

匿名：有助于保护隐私和言论自由（Bodle, 2013），然而网络去抑制效应（Online Disinhibition Effect）（Suler, 2004）也被认为是导致网络欺凌和滥用的因素（O’Leary & Murphy, 2019）。

人脸识别：会导致教育强调人脸特征而去人性化以及突出学生性别和种族等诸多问题（Andrejevic & Selwyn, 2020）。

隐私：“数据收集或聚合、知情同意、数据去身份化、透明性、数据安全、数据解读和数据分类与管理”（Griffiths et al., 2016, p.6）都可能涉及隐私。

考核：学生更喜欢“来自教师或同学而不是计算机的评语”（Roscoe, Wilson, & Johnson, 2017）。

缺乏谨慎：系统不能跟人一样在特殊情况下对规则或正常程序做出例外处理（Passi & Jackson, 2018）。

缺乏异议：以为系统是中立的且不会出错，从而诱使我们把责任、判断和决策这些重要事情交给系统完成（Demiaux & Abdallah, 2017, p.5）。

操纵内容：可能“对隐私、民主和国家安全带来迫在眉睫的挑战”（Chesney & Citron, 2018, p.1760）。

操纵用户：在用户身上做试验，检验什么内容会引起最强烈的反应以便达成不可告人目的（Paul & Posard, 2020）。

（二）分析技术不能有效运行

AI是脆弱的（Liberman, 2019）。如果数据有限或代表性不足，它就无法对环境因素或异常事件作出反应。就学习分析而言，结果可能是系统表现欠佳、教学方法蹩脚、推荐不可信或一事无成（最糟糕的情况），举例如下。

错误：可能包括教学干预没有效果或误导，但却没有人为错误的后果负责（Griffiths et al., 2016）。

数据不可靠：分析技术要求数据可靠（Emory University Libraries, 2019），但是情况经常相反。

一致性失败（Consistency Failure）：在分布式网络上运行的系统，如果网络的一部分有故障，此时可能导致一致性失败（Gilbert & Lynch, 2002）。

偏见：机器学习算法包含隐藏在语言使用模式中根深蒂固的种族和性别歧视（Devlin, 2017）。

误读：如在面试过程中AI根据笔记本电脑位置判断是否有目光交流（Metz, 2020）。

传讹：如科学内容分析（Scientific Content Analysis）的开发者声称它能够发现欺骗性内容，研究结果表明事实正好相反（Armstrong & Sheckler, 2019;

Brandon et al., 2019）。

歪曲：如推荐引擎诱发激进化（Tufekci, 2018）。

蹩脚教学法：“AI可能被用于大规模推行蹩脚教学方法”（Tuomi, 2018）。如分析技术设计欠佳，这可能导致评估错误（Dringus, 2012, p.89）。

（三）利用分析技术作恶

居心叵测者（如黑客）引发了分析技术伦理问题，因为其行为证明人们可能利用这些系统进行破坏，举例如下。

阴谋论者：复制分析技术的方法和传播方式，或利用现成分析技术达成自己的目的（Yeung, 2023）。

跟踪者：利用人脸识别系统跟踪他人可能会被滥用（Shwayder, 2020）。

勾结：同时应用多部分析技术引擎可能导致它们自动出现不良行为。如“应用于重复价格竞争的寡头垄断基础模型的AI（Q-学习算法），尽管没有相互‘通气’但总是会做出超竞争定价（supracompetitive prices）”（Calvano, Calzolari, , Denicolò, & Pastorello, 2020, p.3267）。

（四）分析技术从根本上不可信

诸如累犯、治安、恐袭风险、可能受到伤害的儿童和工作表现等方面的预测会造成社会影响，AI特别不适合用于这些目的（Narayan, 2019）。这方面失败的例子并不少见（Dressel & Farid, 2018）。有观点认为即使分析技术的预测是对的，也不应该用于这些目的，举例如下。

预测性维持治安（Predictive Policing）：如“警官可能因为学生出现在名单上而无意中影响其对案件的处理决定”（Lieberman, 2020）。

种族画像（Racial Profiling）：如“AI系统基于法庭记录和其他数据库评估潜在租客，但是这些数据库自身存在反映系统性种族歧视、性别歧视和（针对残障人士的）体能（或智能）歧视的偏见”（Akselrod, 2021）。

身份图谱（Identity Graphs）：如“第三方根据你的社交和在线临场建立顾客图像，然后把它跟他们的内部数据、信用分数和他们所能找到的其他来源数据合并在一起。这样做合法且合乎伦理吗？”（Hamel, 2016）。

自动化武器，不仅仅指用于实战的武器。如通过学习管理系统防止作弊（Oravec, 2022）、执行版权规定或对使用自动公开披露信息未经许可获取学习资源进行管理或采取报复性措施（如恶意软件或病毒）。

（五）社会和文化问题

这一类不是分析技术自身的问题，而是分析技术如何改变我们的社会、文化和学习方式的问题，包括，举例如下。

不透明性：当AI系统做出涉及个人的决定时，没有遵守告知原则（Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020）。

人际关系疏远：如难见真人（Guillaud, 2020）或面试徒有虚名，甚至没有人看一眼求职者的简历（Keppler, 2020）。

不可解释性：如AI系统对个人生活有重大影响但却无法为其决定提供完整和令人满意的解释（Fjeld et al., 2020）。

缺乏问责：如“应该对系统自动做出的决定解释清楚、给出理由和进行审核”（Rieke, Bogen, & Robinson, 2018）。

社会凝聚力和过滤气泡：这是一个增强和巩固原有模式的循环，把人置于过滤气泡中（Pariser, 2012），随着时间推移他们只看到跟自己观点一致的内容。

反馈效应：指AI对事件的预测增加其发生的可能性。

冷漠：如在凸起减速带等候的送货机器人不让路，导致坐在轮椅上的人没办法过马路（Ackerman, 2019）。

未经同意：如谷歌的“夜莺工程”（Project Nightingale）（Griggs, 2019）和谷歌Classroom因其数据收集做法而引发争议（Singer, 2017）。

监控文化：无所不在的监控成为常态（Schneier, 2020）。

权力和控制权丧失：如“学术研究，从内容到体裁结构都被简化成数据，成为用于生产卖回给学者所在机构的产品的原材料”（Watters, 2019）。

是非感丧失：任由AI决定是非，我们最终将会丧失判断是非的能力（Mitra, 2018）。

所有权丧失：如随着AI生成文本的快速发展，最终AI可能对所生成的每一个有意义的文本主张拥

有版权，人类有可能从根本上被挡在创造内容的门外（Carpenter, 2020）。

不负责任：即自动化自组织系统可能不按照其设计者意图擅自运作（Bostrom & Yubkowsky, 2014）。

赢家通吃（winner-takes-all）：指大公司基于数据的垄断而实现“赢家通吃”的情况（Eckersley, Gillula, & Williams, 2017）。

环境影响：如同其他技术（比如区块链）一样（Hotchkiss, 2019），分析技术和AI同样可能对环境造成危害（Meinecke, 2018）。

安全性：指人们可能用难以被发现的方法非法入侵分析技术系统，如导致无人驾驶汽车把停车标志误读为并入主路标志（Danzig, 2020）。

（六）分析技术伦理的范围

大多数传统分析技术和伦理文献并没有涉及上述问题，包括系统决定我们应该做什么的问题（涉及道义性分析技术）以及社会和文化问题。我们有可能正确使用分析技术但仍然觉得其结果违背我们的道德观或造成社会和文化危害。理解伦理和分析技术可以从伦理原则开始，但不能囿于此。

有研究认为我们对伦理和分析技术已经有了共识，这种观点实则有误导之嫌。如有调查（Fjeld et al., 2020）显示，虽然97%的研究把隐私作为一个原则，但是仔细分析不同研究对隐私的理解便可发现，达成共识的比例远低于97%。其他概念也如此，如问责。这些仅是AI领域的研究。如果不限于这个领域（和不限于技术行业这个背景），我们会看到对伦理更宽泛的理解。我们对不同类型问题的反应因实际情况而异——这点应该很清楚。如果AI因某种原因不能正常运行，我们的反应应该（可能是）寻求避免出现这种情况；如果是被心怀叵测者滥用了，则应该通过寻求法律和立法途径解决。批评者指出的问题并非全都是针对AI的伦理问题，虽然忽视这些问题会有伦理上的影响。

四、我们做出的决定

我们讨论技术能否制造诚实、关怀和信任时，不应该考虑AI或分析技术系统自身能否生成必要的情感以产生这些合乎伦理的好结果，因为这取决于数据输入设计者、商业模式以及其他相关因素组成的更大系统。我们能够用技术制造仇恨，或许我们也能够用技术制造关怀。

技术不是一个决定或（甚至）几个决定的结果。考虑涉及分析技术和AI的所有不同过程以分析我们在设计和使用这些系统时所做出的决定——这一点很重要。我们知道它们是如何制造仇恨和偏见的，同样，我们必须研究如何使它们能够制造关怀。我们必须思考如何做出这些设计和使用方面的决定才合乎伦理。我们虽然知道如果我们缺乏谨慎的态度或故意操纵技术，这些系统便会制造仇恨，但却没有对如何研发合乎伦理的系统给予同等重视。因此，本节的目的是想说明通过了解AI 分析技术的实际机制、实际工作流程和我们实际做出的决定，可以看到我们的努力能够带来伦理上的不同以及这种努力如何能够产生合乎伦理的结果。

（一）学习背景

AI应用于具体环境而非真空中，即受到“教与学现有研究成果”的影响（Gašević et al., 2015）。这限制了我们想做什么、想衡量或预测什么以及谁参与其中，即格雷勒和德拉克斯勒（Greller & Drachsler, 2012）提出的包含六个方面的教学模式框架：能力、限制、方法、目标（区分反思和预测）、数据和利益相关方（Seufert, Meier, Soellner, & Rietsche, 2019）。这其中每一个方面都将直接影响到我们的决定。

所有利益相关方都必须参与决策，否则优先考虑任何一方的意见都可能产生不理想的结果（Jaschik, 2016）。具体说来，AI决策包括“鼓励或要求系统设计者和用户在设计AI用途和管理这些用途时咨询利益相关团体的意见”（Fjeld et al., 2020, p.58）。利益相关方包括学习者、教师、研究人员和教育机构（Khalil & Ebner, 2015），可进一步分为“数据主体”（即分析对象）和“数据客户”（即管理或使用分析技术者）（Jambekar, 2017）。

目标决定利益相关方将用分析技术做什么事情。除了诸如“改进学习”这样含糊其辞的目标外，可能还希望提高学习系统效率、改进系统表现、提升学习过程的透明度和提高学生成绩（Buckingham Shum & Deakin Crick, 2012）。这些反过来又是基于各种指标，包括实际成本、学习投入程度、课程完成率、通过率等。目标还可以指

学习的更大好处，如经济效益或公共利益（Drew, 2018）。

（二）AI运行机制

当代AI建立在人工神经网络基础上。这些网络得先经过“训练”然后才能投入使用（见图1）。

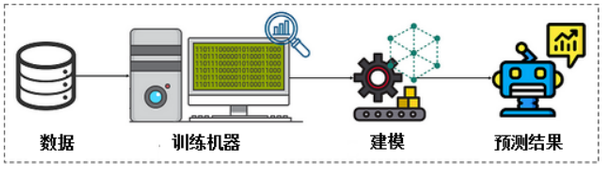

图1 当代AI工作流程（Edureka, 2023）

根据收集到的数据和预期目标，开发者可以选择：①监督学习（有明确的输入和预期目标），②无监督学习（算法自我训练），③强化学习（算法做出决定但可以通过反馈机制纠正决定）。经过训练的网络能够完成各种任务，包括回归、特征检测、聚类和预测。至于软件如何运算和聚类或预测的是什么，则主要取决于学习方式。

一个人工神经网络节点的学习函数决定了每一个单独节点如何接收来自一个或更多节点的输入，应用统计函数计算全部输入，在此基础上决定是否发送信号给其他节点。学习实则是直接或通过学习算法改变影响这个函数的变量（Banoula, 2023）。这些变量包括：阀值（决定一个输入的值是否会触发一个输出的值）、偏差（应用于控制敏感性的输入的负数）、激活值（神经元输入所产生的数值；激活函数是神经元用于根据输入生成其激活值的算法）和权重（改变每一个输入值对神经元的影响程度的倍数）。

神经网络的神经元被组织成一层一层的神经元。在“深度”神经网络中，输入层和输出层中间有一层或多层神经元。不同拓扑的神经网络其功能各异。对神经网络的“训练”是通过利用数据改变连接权重来进行的。例如：一个反向传播（Back Propagation）网络利用来自其输出的反馈，根据“代价函数”提高或降低某些连接的权重。

经过训练之后所得到的神经元和权重的最终配置被称为“模型”。AI的应用便是向一个模型提供新的输入数据以观察其输出的结果。一个模型的创建牵涉诸多方面的决定。虽然大多数决定根本没有明显考虑伦理问题，但是所有决定都可能产生伦理方面的影响。在设计AI应用时，方法的选择（如监督学习或无监督学习）和参数的设置（比如是偏差或阀值）都可能产生伦理影响，虽然这些本身不是伦理决定。

（三）与数据相关的问题

神经网络是利用数据进行“训练”的结果。AI引发的很多伦理问题都与数据有关，因此，数据管理在AI中发挥重要的伦理作用。这方面典型的伦理问题包括数据集成、偏见、标签、特征和建模技术、主观性与风险（Feast, 2019）。其中，数据集成指将不同数据连接起来，此时要比单独一个数据更能揭示隐藏在其中的东西（Cohen et al., 2014）；偏见是由于训练数据集不完整或不准确而产生的；标签指人工标注训练数据以教会模型如何识别数据；特征和建模技术指用作机器学习模型输入的测量参数；主观性指不存在与语境无关的数据，因此数据不能呈现出人们有时所想象的那种纯客观性（Radan, 2019）；风险指数据陈旧过时或代表性不足等（Cohen et al., 2014）。

利益相关各方都有责任发现这些问题并降低其危害（Loshin, 2002）。各方有不同的兴趣和目标，且可能遵循不同的伦理标准。

（四）组织数据

由于AI算法的需要和为了避免出现数据引发的问题，在使用数据之前要对其进行处理，例如：数据清洗（发现、删除和/或替换数据库中不一致或不正确的信息）（Kowalewski, 2020）、数据质量（与其说是源数据的一个特性不如说是数据清洗的一个结果，包括准确性、完整性、一致性、相关性、时效性、有效性和均匀性这些因素）以及分类和命名（在监督学习中包括使用人类可读的符号标注某一数据，相关的操作可以建立在分类、分类法、本体论或自然类型等基础上，而这些是可以在数据清洗之前创建或机器根据经过清洗的数据而生成的）（van Rees, 2008）。

处理数据过程可能会应用到大量标准并涉及许多机制，然而却不存在所谓“正确的”数据标注方法分类。我们的视角、观点或“框架”（frame）决定了我们如何描述数据。数据也可以用算法分类，但是分类算法为数不少。如逻辑回归、朴素贝叶斯、K最邻近、决策树和支持向量机等各自都对数

据分类产生影响（Kumar, 2021）。

（五）算法和拓扑

算法是用输入数据进行训练的，因训练方式的不同而各有区别，如：赫布型学习（Hebbian Learning）、反向传播、数据分组处理方法（Group Method of Data Handling）、竞争学习（Competitive learning）和神经进化（Neuroevolution）。

赫布型学习，经常被归纳为“一起放电的神经元会紧密相连”或“任何两个同时反复活跃的神经元或神经元系统将会‘彼此相连’以至于其中一个的活动会促进另一个的活动”（Hebb, 1949, p.70）；反向传播，即评估错误并通过网络把对错误的纠正反馈回去（Rumelhart, Hinton, & Williams, 1986）；数据分组处理方法，指神经元层的两个输入有很多种组合，这种算法就是要生成适合所有可能组合的神经元，继而挑选出均方误差为最佳的那些神经元（Pandya, Gilbar, & Kim, 2005）；竞争学习，指节点相互竞争对输入数据一个子集作出回应，在这个过程中成为不同类型的输入模式的“特征检测器”（Hassoun, 1995）；神经进化指算法生成神经网络、参数、拓扑和规则的各种方法（Miikkulainen, 2011）。

神经网络也会随着神经元层的组织方式的不同和神经元之间连接的组织方式的不同而异，于是出现不同的网络“拓扑”，包括：前馈神经网络(Feedforward Neural Network)、径向基函数网络（Radial Basis Function Network）、卷积神经网络（Convolutional Neural Network）、循环神经网络（Recurrent Neural Network）、长短期记忆网络（Long Short-Term Memory）、霍普菲尔德网络（Hopfield Network）和吸引子神经网络（Attractor Network）。

前馈神经网络，应用于感知机和多层感知机，数据从输入流向输出等（Rosenblatt, 1958; Upadhyay, 2019）；径向基函数网络即非线性分类方法（Broomhead & Lowe, 1988）；卷积神经网络是节选输入数据的不同部分，通常有一个池化层以缩小矩阵的总体规模；循环神经网络是一个神经元的输出变成该神经元输入的一部分（Donges, 2019）；长短期记忆网络即按照顺序处理数据并在这个过程中保持其隐蔽状态，这样便能够处理数据序列（Hochreiter & Schmidhuber, 1997）；霍普菲尔德网络即记忆可能是一个神经网络的能力最小值，这种神经网络的目的是储存一个或多个模式、根据部分输入找回完整模式（Hopfield, 1982）；吸引子神经网络是一种循环动态神经网络，会随着时间的推移朝着一个稳定模式发展。

算法和拓扑自身没有具体的伦理取向。因此，研发者不能通过“修理”算法使之合乎伦理价值观。然而，算法的选择（或者说设计的选项或算法的挑选）会在很大程度上影响到一个AI系统所能做的事情以及其输出。所以，研发者必须自问：“我的算法是如何与大社会相互影响的？就目前而言，包括是如何处理社会的结构性不平等问题的”（Zimmermann, Rosa, & Kim, 2020）。这是一个AI对齐（AI Alignment）问题，是AI进一步发展的重大挑战（Strickland, 2023）。

（六）模型和理解

应用AI指的是选择经过预训练的模型并应用于具体事情上。这种选择可能对结果产生重大影响。如创建一个模型预测哪些病人应该接受额外照料时，用错算法可能是一个严重问题（Young, 2020）。

至此我们一直用数据、算法和拓扑来阐述AI。在多数情况下，这样做不足以解释AI的运行。解释即是对模型的理解。如检测到一串新单词是一回事，理解其意思是另一回事（Liberman, 2020）。我们对模型的理解是以什么为基础的呢？“使用黑盒子模型使我们难以确定它是如何做决定的”（Dhuri, 2020）。这个过程涉及到的都是数字和统计数据。“海量数据和应用数学取代了可能对世界造成影响的其他任何一种工具”（Anderson, 2008）。数字本身不能够说明一切，而安德森（Anderson, 2008）却全盘否定其他理论和学科。这表明“在很多围绕大数据的辩论中存在一股傲慢的暗流，从而导致其他分析方法很容易遭到排挤”（Boyd & Crawford, 2012, p.666）。

界定一个模型意味着提出一个问题，而问题的选择非常关键（Seufert et al., 2019）：什么问题急需优先处理？结果将得到如何利用？出现不良结果时我们该如何应对？结果将如何测量？模型需要经过“训练”，然而训练则是大量编程的结果。那么，

这个过程使用了严格的编程标准吗？这个程序是否开源？

或许我们可以用稍微不同的方式提出这个问题：AI是以何种方式看世界的？它是如何理解所有这些数据的？理解是一种技能。感知专长（Perceptual Expertise）是“一种针对某个特征或类别的感知辨识或区别的增强能力”（Vance, 2001）。如有些人可能擅长辨识或区分鸟的品种。我们可以把AI看作是擅长辨识的机器。问题是这种能力总是有益无害的吗？我们无法保证感知专长会促进真信念（true beliefs）或知识的增长（Vance, 2021）。感知专长会发现感知到的区别，但是并非所有区别都有意义或有作用，因此可能影响到我们认识或理解某种现象的能力。

（七）测试、应用和评估

软件测试的目的通常是为了确定其程序会产生预期的结果。每一个阶段都要经过测试，包括原始请求（确保收集的数据正确、向系统发出的请求正确、请求发送方式正确等）、数据测试（检验有效性、可靠性、多样性、一致性等）和应用（安全性、性能、可用性和故障转移）。但是，任何模型最终都必须在现实世界的应用中进行测试，当然必须做足合适的保护和预防措施（Cohen et al., 2014）。

一个AI应用必须先介绍给目标使用者并被他们所采用，否则不能发挥作用。这是加拿大卫生研究院（Canadian Institutes of Health Research）在2000年提出的“知识转化”（Knowledge Translation）的范围，该术语指“知识的交流、整合和合乎伦理的应用”。“知识动员”（Knowledge Mobilisation）则是一个较新的术语，是“指与研究成果的生产和使用相关的诸多活动的一个统称，包括知识的整合、传播、转化、交流以及研究者和知识使用者共同创建或生产知识”（Wilsdon et al., 2015）。

AI应用不能不考虑时机问题。“我们不能忽视数据和分析技术中人的元素。仅做到准确分析、预测和可视化还不够。大学的师生都必须具备数据素养以便能够理解和利用数据。只有师生能够理解呈现给他们的东西以及知道他们据此可以做什么和如何做，才有可能采取合适和有效的干预措施”（Clay, 2020）。

至于评估，其目的不是检查AI或分析技术的应用是否正常，而是确认AI的使用是否产生令人满意的结果。但是，何谓令人满意的结果在很大程度上因人而异。因此，评估必须置于更大背景下进行，包括与AI的设计和开发毫无关系的因素。如我们不仅仅根据是否“促进学习”来评估学习分析技术，而且还会考虑诸如“支持联合国可持续发展目标”“提高组织的效率”或“为股东创造更多价值”等问题。

五、伦理准则

针对AI引发的伦理问题，普遍认为应该制定伦理准则，即确定哪些是有争议的伦理问题并一一提出对策。遵守伦理准则便能够规范涉及AI的伦理行为（简称“伦理AI”）。

（一）原则

AI伦理准则的一个主要特点是包含一套倡导者共同认可或应该被广泛应用于整个AI领域的伦理原则。这种主张通常没有言明，虽然言明的情况也不少。从历史上看，诸如人权、自由、不伤害、正义和公平等自由民主价值观是被普遍接受或主流的伦理价值观。如有研究发现各种伦理原则高度重叠并针对伦理AI提出了一个包含五条核心原则的框架（Floridi & Cowls, 2019）（见图2）。

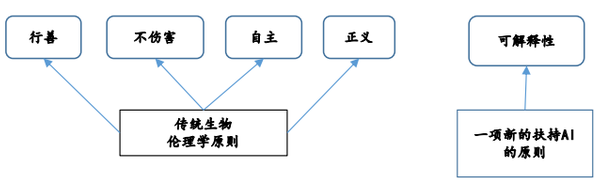

图2 AI原则的伦理框架（(Floridi & Cowls, 2019)

但是弗洛里迪等（Floridi et al., 2018）原则的提出者明显具有同质性，包括阿希洛马会议（Asilomar Conference）出席者、欧洲委员会成员、电气与电子工程师协会（IEEE）会员和AI利益相关者。这些人的共同点是对AI、立法和政策感兴趣，因此认为必须制定伦理准则管理AI的应用。这种观点是可以理解的，它源自各行业都有伦理准则这个传统，也有助于表达一套共同的价值观。如梅特卡夫指出：“我们发现很多领域当代伦理准则的核心是一些原则，包括对人的尊重（自主、隐私、知情

同意）、平衡对个人的风险和对社会的贡献、谨慎挑选参加者、独立审核研究计划、成立专业人士自我管理的组织和根据遵守伦理标准情况拨款”（Metcalf, 2014, p.2）。应该指出，梅特卡夫的这一套原则与弗洛里迪和考尔斯（Floridi & Cowls, 2019）的原则有很大区别。

进一步分析还显示，这些原则即使貌似有广泛共识之处，细节上也存在很多不同。有一项研究（Fjeld et al., 2020）断言其所述评的那些伦理准则包含共同的原则，例如：研究者指出约70%的准则对“透明性”和“可解释性”的理解相同，但是如果进一步细分，达成共识的比例远不如人意。具体说来，“开源数据和算法”“知情权”“告知互动的对象是AI”“开放政府采购”“告知涉及个人的决定是AI做出的”和“日常报告”等的共识比例分别只有28%、11%、25%、3%、19%和17%。值得注意的是，这些伦理准则的制定者具有相对同质性。图3能很好地说明这种“各执一词”的情况。

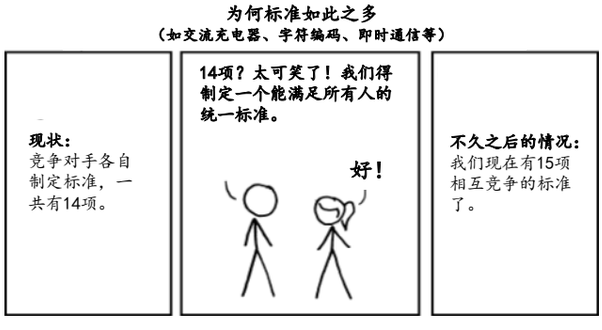

图3 为何标准如此之多（Cummingham, 2008）

（二）价值观

我们可以说“一个人的职业责任源于其行业和行业准则、传统、社会期望、契约、法律和一般道德规则”（Weil, 2008），但是这种观点解决不了什么问题。我们在分析伦理准则文献时发现它们是建立在众多价值观或理由的基础上的。如很多准则把普适性作为伦理和道德原则的一个依据，例如：《心理学家伦理原则普世宣言》（Universal Declaration of Ethical Principles for Psychologists）指出其伦理原则建立在共同的人类价值观基础上（IUPSYS, 2008）。与普适性相关但不同的是基本人权，例如：人工智能高级专家组（High-Level Expert Group on Artificial Intelligence）引用了四条伦理原则，这些原则是“从基本人权发展而来的，必须得到尊重以确保AI系统以可信的方式研发、部署和使用”（AI HLEG, 2019）。

然而，我们可能还要多从务实角度考虑。讨论AI伦理经常要考虑风险与回报的平衡，如《民享AI》（AI4People）框架指出：“一个伦理AI框架必须旨在使机会最大化而风险最小化”（Floridi et al., 2018, p.7）。这在很大程度上是后果论（Consequentialist）的方法，因此导致每一个AI应用采用不同的计算方法。此外，这种方法还要求我们知道实际的结果是什么，或者说伦理并非真的是关乎平衡各种相互竞争的利益，而是建立在最大化回报的基础上。一位来自加拿大渥太华的信息和隐私专员坚称“采用正和方法（positive-sum approach）设计规范国家监控的监管框架能够避免假二分法和不必要的权衡取舍”（Cavoukian, 2013）。

过去3 000多年来人类一直在讨论伦理问题，但是尚未对伦理的基础、本质或原则达成一致意见。

六、解读伦理的方式

（一）美德和品格

从这个方面讲，伦理首先是研究人的美德的学问。美德可能是柏拉图式理想或体现在一个人在社会中的行为举止上。美德通常包含高尚的特征或特质，如诚实、节俭、虔诚、谦逊、关心、勇气等。这样的美德，亚里士多德列举了12种，而在儒家看来美德包括仁、义、礼、智、信（Wahing, 2021）。但是，定义美德的不是这些特征，而是集这些特征于一身的品格。一个拥有美德的人为人处世品行端正，而美德本身则是一个为人处世品行端正的人必不可少的道德特征。

美德理论的隐晦既是其优点也是其主要缺陷。如尼采曾经提出“‘超人’可能拥有哪种美德”这个问题（Nietzsche, 1999）。

（二）义务

对康德及其追随者而言，伦理建立在义务的基础上。康德的伦理基于两大原则：一是绝对命令（categorical imperative）原则，即伦理原则应该是普适性的，所以我们应该关心的是：如果大家都这样做，情况会如何？二是人本身是有价值的，我们应该把人当成“目的”而非“手段”。人不是被使

用的物品；作为理性和通人情，尤为重要的是知晓伦理的生物，人有内在价值和地位。

这些原则要求我们履行一个义务，即我们的行为必须对社会和社会中的个人都有好处。这在直觉上很诱人；“能力越大，责任越大”。只拥有美德还不够，我们必须践行美德。然而，什么样的美德是普适性的？我们又是如何把彼此看作是目的或手段的？人们常常引用康德伦理学为自然主义辩护，即我们应该避免非自然的行为。但是或许人的身体能够做的一切事情都是自然的，因此以非自然为理由反对人的行为难以令人信服。的确，如果描述准确，任何行为都可能被视为普适性的。

（三）后果论

后果论（Consequentialism）是一个概括性术语，包含众多伦理学理论，从“不为害”到“只要目的正当，可以不择手段”，不一而足。“后果”即行动或状况的结果、影响或重要性。在后果论看来，合乎伦理的行动或原则应该根据其后果进行评估。

不同的后果论流派对何谓可取的后果有不同解释。对于个人而言，幸福是可取的，幸福可以被描述为感到快乐、没有痛苦。但是，侧重点可能各异：享乐主义者只在乎生理愉悦，或如同米尔所建议的，我们不妨追求知识和启智带来的“高层次快乐”（Mill, 1879）。伊壁鸠鲁学派信奉者（Epicurean）可能认为快乐即不用遭罪，教导人们应该追求一种“毫无纷扰”（ataraxia）的状态。避免痛苦可能如同佛教所言是一个态度问题：人们因处于无法控制的境地而经历难以忍受的痛苦，世界一直在变化，如果人们要追求永恒，必定遭受痛苦。后果可能针对个人也可能针对集体。墨子曰：故古者圣王，明天鬼之所欲，而避天鬼之所憎，以求兴天下之利，除天下之害。这是一种社会后果论（Harris, 2017）。

同样的，不同社会对可取的后果有不同理解，如有的地方，指的是生命、自由和追求幸福，有的地方指自由、平等和友爱，而有的地方则是和平、秩序和良好管治。

（四）社会契约

在本文，社会契约的核心是把伦理视为一个社区的共识的结果。社会契约伦理的主要组成部分是：①达成一致意见的过程或方法；②确定一致意见的内容；③遵守业已达成的一致意见。不同社会契约伦理理论对这三个部分有不同的阐释。如《正义论》（A Theory of Justice）的作者罗尔斯提出假设性“原初状态”（original position）概念，指参加者经过“无知之幕”（veil of ignorance）的甄别之后参与社会契约的协商，这就是正义即公平的理论（Rawls, 1999）。特别是每个人享有平等的基本自由的权利，所以他们首先创造机会公平平等的条件，其次是使社会最弱势的成员成为最大受益者（Rawls, 1999）。

那么，是什么促使一个人接受社会契约呢？因为不接受的话，情况可能更加糟糕，我们会过上“孤独、贫困、污秽、野蛮而又短暂的”生活（Hobbes, 1994）。或许我们也可能把这些权利看作是与生俱来的，如同卢梭所言“人人生而自由”（Rousseau, 2004）。或许是承认我们本质上是社会性的，因此要求有支持这种社会性的条件。不管是以宗教教义形式还是政治宪法、宣言和协定的形式，历史上出现了各种各样的社会契约，有多少种社会契约可能就有多少种接受契约的诱因。

（五）元伦理

伦理的解读方式不限于上述四种，但是它们自身以及彼此之间意见不统一，这说明有必要在更大的背景下讨论伦理的基础这个问题，即为什么说某一伦理的立场是对的？

这方面存在各种可能性。我们在回顾康德的理论时发现普适性这个概念，即一条伦理原则必须如同自然法则一样可被普遍应用，也有人提出自然本身就是伦理的基础。有什么比一个人自己的身体、感官和情感更加自然的呢？或者道德最能表达我们的认知能力以及理性和启蒙，或许伦理更像我们的感官，是非认知性的，或许伦理如同科学，帮助我们发现对与错，也可能对与错的发明是为了服务某些其他目的。伦理难道是赖以做决定的依据或事后提出来的吗？

（六）告别传统伦理

现在我们必须放弃传统伦理的思想。显然，我们有伦理，但是不知道为什么有伦理。在这点上，人的认知与机器学习一样神秘莫测。有研究者指出：“在‘我们不知道机器如何学习’的叫喊声中

我们也听到这些模型的确有作用……我们在与机器学习模型打交道的过程中看到存在普遍性的现象、规律或原则，但我们认为这些东西不足以帮助我们理解人类这个复杂的世界”（Weinberger, 2021a）。

简单的原则无法应对复杂的思想、想法或问题。我们现在“必须处理的是相互依存的问题，经历非线性且常常不可预测的变化过程，牵涉到各种各样的利益相关者”（Jones, 2011）。首先，解决复杂问题的能力经常埋藏在参与者之中。其次，复杂问题难以预测，很多社会、政治和经济问题难以详细预测。再次，复杂问题经常涉及互相冲突的目标。

伦理不是抽象的。虽然我们可以根据抽象价值观或估算归纳普适性原则，但是我们不应该这样做。传统伦理是一种冰冷、静态、受制于规则、死板、严苛的伦理，如同寒冷的西伯利亚般令人发怵，而本文提出的是一种温暖、动态、不是非对即错、灵活、宽宏的新伦理，能够给人快乐（Rushkoff, 1994, p.180-182）。只有倡导和恪守的AI伦理是一种温暖、动态、不是非对即错、灵活、宽宏的伦理，才能够用AI创造美好。

七、注意义务

伦理常常被视为事关论证。如果我们提供正确的理由，人们就会明白解决某个困境的伦理方法是什么，然后按照这个方法去做事。论证法建立在诺齐克的“威迫哲学”（coercive philosophy）的基础上，即“无法反驳的论证非常强有力，也是最好的，论证迫使你得出结论”（Nozick, 1981, p.4）。

然而，事实常常与此相反。论证不能使任何人得出结论，如果说有什么作用的话，那就是迫使人们提出反对意见、更加坚定自己的信念。“人们赞同的那些道德原则涉及他们的生活期望、家庭角色和社会地位。很多研究表明，人的道德判断受其年龄、性别、家长地位、教育、多元文化经历、战争经历、家庭生活或宗教信仰等影响”（Ellemers, van der Toorn, Paunov, & van Leeuwen, 2019, p.351-352）。

辩论常常只是使我们更加坚信自己的立场，至少从我们的角度讲自己的立场没有错。相反，可以说对很多人而言，以道德服人和道德判断关乎的是我们与我们对其负有照顾责任的那些人的关系，如母亲对小孩的照顾（Weinberger, 2021b）。这就是“注意义务”（Duty of Care）的基础。

（一）注意作为一个法律概念

注意义务作为一个法律概念可能源于阿特金大法官（Lord Atkin）处理的一宗案件——一瓶姜汁啤酒里面出现一只腐烂的蜗牛。法庭面对的问题是负责啤酒生产和装瓶的人是否对消费者承担责任，尤其是当消费者出现不良反应的时候。对此，法庭未能达成一致意见，因此做出了分歧裁决（split decision）。如果一个人在实施可以预见可能会给他人造成伤害的行为时，这个人有遵守合理注意的标准的法定义务。当然，这个例子不同于母亲对小孩的照顾，但是二者有很多相同之处。照顾者和被照顾者之间的关系是不对称的，这种不对称性导致一方要对另一方负责。有时，这种责任以法律形式规范下来，有时则似乎更是一种生理必然性（biological imperative）。

从法律意义上讲，一个人有责任或法定义务避免可以合理预期的会给他人带来伤害的行为或不作为。如果存在某种特殊关系，如教师与学生，这种义务就变得更加具体和迫切了。

但是，作为伦理的一个标准，上述的这个去人性化的功能性定义显然不适合作为注意义务的定义。因为新技术具备新能供性，如AI和分析技术，把注意义务作为一个法律概念理解会有一些重大局限，尤其是这些法律原则没有考虑师生之间、学生与学生之间以及社会成员之间非常真实、非常重要的关系。

（二）注意作为一个伦理概念

维基百科指出注意伦理（Ethics of Care）是一个规范性伦理理论，根据这个理论，道德行为以人际关系和关怀或仁爱为中心并视它们为美德。这个定义显然是基于其他伦理理论的术语和分类法，不是从注意本身的角度出发。但重要的是它不再把没有人情味的人际关系当成注意的核心，换言之，我们应该尊重和关心与我们交往的人，哪怕彼此之间没有私交。“关怀、照顾、谨慎和被照顾是可被感受到、多维度、使人更加自信、令人担忧和随时出现的，不是单一和静态的”（Motta & Bennett, 2018, p.640）。如服务提供商与顾客之间的关系应该是一种真诚交流和关注另一方的需求的关系，不是死板

和程序性的关系。

注意这个概念要比注意伦理早出现。它被视为一个过程、一种与人交往的方式且会发展，犹如友谊只有在互信的基础上才能建立起来并不断加深、促使关系发生质的变化一样。关心一个人的最大意义在于帮助其成长和实现自我（Mayeroff, 1971）。这里讲的是共同成长、共同发展和注意本身的关系性。注意常常成为我们的其他价值观和活动的中心。

注意作为一个伦理概念源于女性主义认识论和女性主义伦理学，因此切记不能脱离具体背景去理解注意。把研究注意伦理的学者们看成是各自在提出一个自己的、独立的理论并且围绕伦理应该是什么展开争论，我认为这是一种误读。毫无疑问，学者们都有自己的观点，而且这些观点是非常鲜明的。但是把注意伦理学当成一个统一的整体考虑，认为学者们只是在突出这个整体的不同方面，而不是在争论谁的注意伦理观更正确，这样会更合适。

（三）注意与关系

注意作为一个理论的核心假设是：人有不同程度的依赖性和相互依赖性。一个人选择的结果会影响到其他人。把一个弱势人士当成注意的被动接受者是不够的。相反，注意建立在关系的基础上，建立在照顾者和被照顾者之间相互交换的基础上。重要的是二者之间关系的具体体现，不是各方利益的具体体现。这些都不能被当成抽象概念进行描述。

注意这个概念不是从人性或美德中演绎而来的，而是扎根于心声表达和关系之中并坚持让每个人都能够发声、按照自己的主张和根据自己的意愿被认真倾听、得到重视的一种伦理。此外，需要得到回应并发展关系（Gilligan, 1982）。

因此，注意的基础是真正考虑受到决定影响的人的真实声音，不管这些决定涉及什么，包括被照顾者的心声，但不一定局限于这些人。注意不是建立在我们如何想的基础上，而是基于人们表达出来的需要，基于与被照顾者（可能还包括其他人）的实际交流。他们只有通过交流才能表达需要，其他人无法讲清楚这些人需要什么。

注意伦理的前提是作为人类我们从本质上讲是关系性的。人类的环境就是一个相互连通或相互依存的环境。因此，道德是源于反映我们如何与他人相处的一种心理逻辑，不是建立在理性论证的估算或逻辑的基础上（Gilligan, 1982）。因此，注意被视为一个女性主义理论，至少在一定程度上是因为其倡导者相信女性更可能基于关心、包容和私人关系做决定，基于她们生儿育女的经验做决定，而不是基于正义这种更加抽象和冷漠的概念做决定。

注意变成一种义务，因为关系意味着一方对另一方作出反应的重要性和紧迫性。这个义务不是关乎做一个更好的人或对另一方进行道德评判。因此，不要局限于从字面上理解“注意”，它也是一个富有情感或能激发积极性的概念，伦理指的就是我们在这个意义上做出的反应。因此，讨论照顾者和被照顾者之间的社会契约甚至可以说没有意义，同样，从后果论的角度讨论伦理也是没有意义的。

（四）注意教学法

如果我们把注意伦理应用到教育上，则不但能够建立一种独特的“注意教学法”（Pedagogy of Care），而且还会重塑分析技术在学习中的角色。传统上学习分析旨在帮助教师设计学生学习体验，而注意教学法则旨在“确保学生有足够的知识和工具为自己做出明智的选择并知道能够在多大程度上设计自己的路径”（Bali, 2020）。“从教学上讲注意即承认学生的复杂的、富有创造性的能量、愿望和经历是知晓可能性（knowing-possibility）之源”（Motta & Bennett, 2018, p. 637），即教师不是知识的唯一来源。如果我们赋权给学生，允许他们根据自己的兴趣和意愿学习，他们就会有“创造力”，这种创造力是知识的源泉（Bennett et al., 2018）。教师变成给予鼓励和信任的角色，相信学生能够从自己的经历中学到知识。

（五）道德情操

女性主义伦理理论给唯理性的思想体系当头一棒，后者可能从根本上忽视了知识建构本质上是个人的，有时也是基于性别的这个性质（Dunn & Burton, 2013），换言之，伦理知识从显性知识范畴中被转移到波兰尼所言的隐性知识范畴中（Polanyi, 1966）。我们的伦理行为既不是推断也不是推测行为，而是如同马歇尔的“伦理警报”（ethics alarms）一样（Marshall, 2023），是我们的感觉。“于是，情感和情感的体现对知识建构和知晓（知识的）主体至关重要，尤其是有关教育和体现包容

或排斥、正义或非正义的教学法的知识建构而言更是如此”（Motta & Bennett, 2018, p.634）。伦理与其说是一种认知不如说是一种感觉，我们称之为一种明辨是非的能力或休谟所言的道德情操（Hume, 2007）。

进一步从明辨是非能力这一点展开，我们可以借鉴拉德克里夫的观点，即道德的明辨依赖我们的经验、情操或情感（Radcliffe, 1994）。这不是先天论或自然道德论，不是说天生懂得道德是什么，不是笛卡尔式的确定性（Cartesian certainty）（例如：“我认为……，所以我……，所以我是道德的”）。我们能够学习伦理，但是我们是在感觉或体验明辨是非中学习伦理的，不是从完美的一般原则中学习伦理的。必须指出，这与道德直觉不同。人们在提到注意伦理时可能把它与他们凭直觉在谈论的事情等同起来，如女性直觉。但是注意伦理不是这样，而更像是一种情操或情感，更接近于一种平衡感——我们失去平衡时的那种感觉不是直觉。注意伦理常常是一种亚符号（或不可言喻）层次上的体验；伦理关乎的不是理性而是同情。我们在某个场合的反应取决于我们的伦理背景，是多种因素同时作用的结果，而不是因为制定了那些博眼球的大原则的结果。

道德感觉不是诸如愤怒感、恐惧感或欲望感这一类的情感，而是更加细腻的一种感觉。同样的，我们也可以认为这些感觉“影响行为准则的文化演进，如集体共享自主行为准则，部分原因是这些准则唤起相互尊重的道德感并因此在文化传播中受到青睐”（Kumar & Campbell, 2023）。

八、学习分析中的伦理实践

综上（从AI用途到注意义务）所述，AI谈不上“应用”伦理，没有哪一个人或集体要为伦理结果“承担责任”，对于什么是“伦理AI”也没有专门说法。因此，认为我们能够采取干预措施以研制出“合乎伦理的”或“可以解释的”AI似乎有误导之嫌。但是在具体情况下我们可以谈论AI伦理，诉诸我们自己的道德意识判断AI的某一项应用正确与否。在实践中我们对一切事情的伦理的看法，会因我们对其潜在后果的态度的不同而异（图4）。

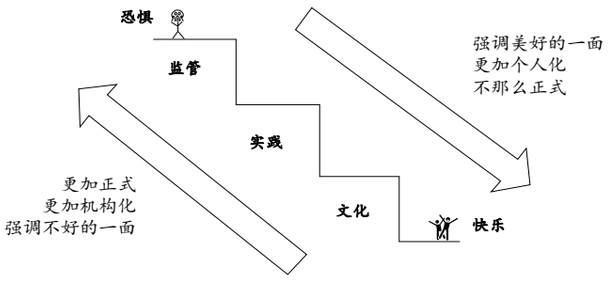

图4 从监管到文化

“今天新兴技术以及围绕它们的讨论，如同神话般弥漫着一种难以理喻的气氛——从根本上无法理解或审查预测性工具的‘决策’，给人一种不容置疑的感觉——不敢质疑这些技术是否有悖于我们人类的伦理框架”（Tsui, 2023）。改变AI的方向，这样才能“驯服”这种神秘技术——这点很重要。因此，应该明确我们希望AI往哪个方向发展、AI应该服务于哪个方向（Tsui, 2023）。通过强调AI重在回应全球绝大多数人的需要和使AI按照指定方式服务这些目的，我们能够控制它的神秘性（Tsui, 2023）。

（一）监管和管治

监管往往强调AI引发的最直接相关的伦理问题，如公平、透明和隐私。监管者关心的是可解释性和可理解性，如算法决定的“解释权”。监管经常是依靠伦理审查，即检查算法的输入和输出中是否包含偏见和有害内容的机制（Cath, 2018）。我们在之前“模型和理解”一节讨论了对AI进行解释的内在问题。从管治的角度看，“可解释性”是建立在更加正式的标准的基础上的。如英国信息专员办公室（Information Commissioner’s Office）和艾伦·图灵研究院（Alan Turing Institute）指出可解释性的六种主要类型，包括理据（一个决定背后的理由）、责任（谁制造AI系统、如何组织人力对系统进行审查）、数据（输入模型的是什么数据、如何使用数据）、公平（如何知道AI没有偏见和公平对待每一个人）、安全和性能（如何确保准确性、可靠性、安全性和稳健性）和对影响的解释（效果和决定是如何受到监控的）（Schildkraut, 2021, p.9）。这些不是科学意义上而是法律和政治意义上的解释，旨在寻找所采取的行动的具体证据。

人们经常采用基于风险的方法。如《欧盟人工智能法案》禁止某些AI实践，包括阈下意识（subliminal）、利用人的弱点、社会信用等级、实时生物指标（取决于具体情况）等的技术（EU AI Act, 2021）。其他司法辖区也以存在滥用风险为由禁止人脸识别技术的使用。欧盟《通用数据保护条例》（General Data Protection Regulation）包含数据权利，而知识产权法则对AI生成的内容的“著作权”和“所有权”进行定义（目前尚无这方面的法律，已有一些裁决确定AI生成的内容没有版权）。一些条例也涉及如何处理与AI有关的民事侵权行为，如制造和设计上的缺陷或未能告知风险（Hodgett, Liu, & Ip, 2023）。

尤为值得关注的是涉及人权方面的管治。这方面“不只是提出如何‘不为害’或‘遵守伦理’的标准，而且会有助于使机构对这些标准负起责任”（Biddle & Zhang, 2020）。如哈尔伯塔尔定义了三种违反人权的行为：“被置于无助、微不足道的境地，失去代表自己的自主权；视人为可交换的物品，只不过是达成目的的手段；把一个人当成是多余的，不承认其贡献、抱负和潜能”，所以要从人权角度出发设计AI（Halbertal, 2015, as cited in Aizenberg & van den Hoven, 2020）。然而，人权的定义可能会过于西方化、过于强调个人主义，此外涵盖范围太小、内容太抽象，不能成为稳健管治AI的基础。

就AI伦理而言，监管法有本质上的局限。符合法律的经常不合乎伦理，监管受到古德哈特定律（Goodhart’s law）影响，即当一个指标成为优化的目标（并被篡改）时，它不再是一个合法的指标。如设立捕捉眼镜蛇奖金旨在减少眼镜蛇的数量。一开始这个奖赏制度运行良好，但是，当人们开始饲养眼镜蛇以领取奖金时，眼镜蛇的数量反而增加了（Treviranus, 2018, p.32）。同样的情况还包括专门研究如何绕过AI监管条例，根本不考虑AI伦理（Rodrigues, 2020）。

（二）伦理实践框架

“我们应该把我们知道有作用的伦理嵌入机器中（即设计伦理），还是应该把希望寄托于通过机器社会（society of machines）的协商取得伦理共识这种方法？”（Nallur & Collier, 2019, p.534），即合乎伦理的行为准则不是事先设定的，而是通过多智能体的软件系统在具体情况下涌现的。第一种方法似乎是最佳选择，但是在一个不但伦理观点众多而且相互影响的AI系统众多的世界里（即机器社会），伦理是不可能被设计的。因此，考虑呈现不同伦理视角的AI系统如何才能对伦理问题达成一致意见这种方法更为可取。换言之，采用的是伦理实践或框架，不是具体的准则或原则。

规则和原则一旦应用于实践就会打折扣。因此，很多旨在支持行业行为的方法聚焦共同实践而非原则，因为虽然实际结果和最佳决定不可能预测，但是在某一种具体情况下遵循一条标准就会产生最佳结果（Courtney, Lovallo, & Clarke, 2013）。

有些伦理实践框架可能包含伦理管理框架、数据管治框架、IT管治框架或人权框架。每一个框架专门处理AI系统和实践的相应方面的问题。如一个伦理管理框架可能会建议应该采取的一系列步骤。如根据圣塔克拉拉大学（Santa Clara University）马库拉应用伦理学中心（Markkula Center for Applied Ethics）的框架（Kwan, Mclean, & Raicu, 2021），我们在发现一个伦理问题之后，要了解相关情况并评估其他方法，然后做出决定并反思结果。采用快速结果评估法（Rapid Outcome Mapping Approach）的“支持高等教育融合学习分析”（Supporting Higher Education to Integrate Learning Analytics）框架则专门应用于基于科学证据的决策（Young et al., 2014），它建议发现问题、提出对策以及制定监控规划和学习计划这些步骤。

这些框架对实际或可能出现的问题作出反应，所以不是主动预防性的。加拿大隐私专员公署则主张制定与AI相应的法律，包括一个事前监管框架，允许在不违反人权的情况下可将私人信息用于新的目的，同时制定专门针对自动化决策的条款，要求企业负起责任（Office of Privacy Commissioner of Canada, 2020）。相比之下，像Digital Catapult这家机构的框架则更像是一份清单，其主要目的是帮助AI公司设计和部署伦理AI产品，包含七个方面的内容：①清楚说明产品或服务的好处；②知道并管理风险；③以负责任的方式使用数据；④值得信赖；⑤鼓励多样性、平等和包容；⑥持开放和易懂的交流态度；⑦考虑自身商业模式（de Bruijn et al.,

2019）。虽然伦理框架旨在支持制定相关规定，但它们经常不发挥监管作用。这些框架界定哪些实践可以被视为行业或机构的实践，采用各种机制管治实践，比如决策树、清单、框架和过程。这些机制没有对伦理实践进行定义，但是都有助于鼓励合乎伦理的实践或防止出现有违伦理的实践。

（三）伦理社区

从AI伦理的角度讲，管治和伦理框架还不够，因为这些是针对机构的，不是为了服务大社会。它们建立在共识和共同假定的基础上，实际上大多数是没有在伦理理论指导下制定出来的。没有遵循这些框架会被视为违反法律或机构的规定，但本质上不违背伦理（除非把“伦理”严格定义为一个法律概念）。因此，遵守和执行这些框架的情况一直不尽如人意，人们试图通过诸如商业风险管理、培训和发展计划以及高调表态的原则和价值观等机制处理这些问题，但是收效甚微（Blackman, 2020）。

有时，这些框架是从公民素养的角度制定的，虽然人们对公民素养的认识存在很多差异，而且似乎经常以某一方式解读伦理，从洛克自由主义（Lockean liberalism）——“追求美好生活和不受政府不合理的干预”、共和主义（republicanism）——公民广泛参与是一种社区义务（Mossberger, Tolbert, & McNea, 2007, p.6-7）到所谓“数字公民”——“信奉理性主义和尊重公民自由和自由市场经济”（Katz, 1997）等视角解读伦理。但是，数字公民素养这个概念经常脱离实际生活（TeachThought, 2019）。

伦理问题最终是文化问题，不是管治或框架的问题。所谓“文化竞争战略”，这点已经在商界得到承认。或者说，伦理问题是某一种实践或方法的“合法性”问题（Schintler, McNeely, & Witte, 2023）。所以，AI伦理是AI的研发、生产和使用的参与者的文化问题，重要的是如何处理AI的伦理文化。

数字公民素养的一个目标是培养伦理社区的概念。公民素养涉及的范围太小，尽管它源自全球伦理这些自上而下的原则。有研究者主张“进一步缩小范围，聚焦网上尊重他人的行为和网上公民参与……控制其他变量之后，网上尊重他人行为和公民参与与在线骚扰行为存在负相关的关系，而与促使出现有作用的旁观者行为存在正相关关系”（Jones & Mitchel, 2016, p.2063）。

从AI系统的角度界定合乎伦理的实践，把AI当成社区一个遵守伦理的成员——这是有益的。虽然我们往往从美德和义务的角度阐述伦理AI，但是归根结底还是在于AI和人类能在多大程度上顺利“相处”。我们可能认为跟我们有关的只是AI的输出，但其实我们在输入阶段就已经与AI发生关系了，即用数据训练AI。如同人类一样，AI从所处的文化中学习和模仿这种文化。

另一种伦理社区模型是基于参与的概念。有研究者从数字素养出发阐述这种关系，指出“媒体制作者和社区参与者这些角色越来越大众化，传统上培养年轻人扮演这些角色的职业培训和社会化不奏效了”，因此建议培训新技能，包括分布式认知、集体智慧、发展人际关系和协商等方面的技能（Jenkins, 2006, p.3）。这个模型可能把我们与AI的相互作用当成社区的一部分并以此指导AI设计（Amershi et al., 2019），也包括有效使用AI所需的技能，从AI素养（Long & Magerko, 2020）到提示工程（Wang et al., 2023）等。

（四）伦理和文化

如上所述，伦理AI的出现取决于伦理文化。那么，什么是文化或伦理文化？同样，这个问题也存在不同观点。我们可以把文化看成是“某一特定人群的特征和知识，包括语言、宗教、饮食、社会习惯、音乐和艺术”（Pappas & McKelvie, 2022）或者是“通过社会化学习的共同行为和交往模式、认知构念和理解”（Damen, 1987, p.367）或者是“对思维的集体编程以把一类人的成员跟另一类人的成员区别开来”（Hofstede, 1984, p.51）。

无论从哪个角度讲，文化是在一个由交流、交往、行为和传统组成的关系空间里形成或建构的，可以对其进行静态定义（如前文所述），也可以进行动态定义。如有研究者把文化比喻为皮氏培养皿里面所发生的事情，即培养、养殖。不管是静态的还是动态的，文化是通过社区中的交往和参与形成与发展的。与注意伦理一样，文化不是建立在原则或普遍性东西的基础上，而是建立在个人和集体行为的基础上。有研究者认为这些行为有行动主义的形式，类似“亲密式叙事”（close narration）

中“悄然无声的关怀举动”（quiet acts of caring）（Haffey, 2023），是一场运动，旨在实现“每一个人都能够根据自己的愿望、兴趣和爱好自由通行并完全享有生活必需品的社会愿景”，不是为了“提升境界或争取得到承认或提高公民素养”（Hartman, 2018, p.471）。

这可以被归纳为“连接行动”（connective action）（Bennett & Segerberg, 2013）的一种形式，即以“我们可能利用平台互相照顾而不是退出平台”（Singh, 2020, p.147）的“平台倾向”（platform inclination）取代“对抗”或“反对”，以“制造希望”取代“在天空中寻找希望”；区分“把注意作为一种表演”和“以行动表达注意”，区分连接行动与“传播劳动”（communicative labour）（即要重视利用社交媒体把个体组织起来而不仅仅是传播更能博眼球的抵抗）（Singh, 2020）。“参与式文化把素养的重点从个人表达转向社区参与”（Jenkins et al., 2009, p.6）。同样的，它也能够重塑伦理的重点。这样的一种文化将鼓励艺术表达和公民参与、非正式学习以及社交联系。

（五）和谐伦理

维基百科指出，“伊里奇（在《陶然自得的工具》）概括了他此前应用于教育领域的那些主题：专业知识的制度化、技术官僚精英在工业社会的主导角色和需要发展新工具让普通民众夺回实践性知识”（Wikipedia, 2023）。这可被视为代表基于社区的实践培育伦理文化的一种方法。陶然自得算不上是一个伦理理论。牛津学习者辞典将其定义为：“欢乐友好的氛围或性格”（Oxford Learner Dictionaries, 2023），即寻找生活中的乐趣并与社区其他成员分享这种快乐。陶然自得不是泾渭分明的事情且被视为小事情，与严苛的纪律和原则形成鲜明对比。它犹如两个人在马路上擦身而过时给彼此留下的空隙以确保没有人会被挤到边缘，关乎的是体贴和合情合理，是我们对彼此应有的小礼仪，以示尊重每个人的情感和喜好，是我们做小事情的时候所表现出来的周到（Brody, 2019; Sheather, 2020）。

传统伦理在很多方面都是针对自己的，基于某种理由规定应该做什么事情。但是，本文提出的和谐伦理（Ethics of Harmony）如同注意伦理一样是为了他人的。因此，它从根本上讲也是一种开放性伦理，是一种注意、倾听和理解其他人缺乏什么、需要什么和能够提供什么的伦理。它是开放的，这种开放不是一个要求（如开源软件或开放内容的“开放”），而是一个通过分享为公共利益做出贡献的机会。“开放即乐于接受他人，包容而非排他。它欢迎多样性，重视（不只是‘容忍’）其他人并试图发现他人的天赋和才能”（Aerisman, 1999）。

重视建立关系和多样性可以被视为是对他人开放、注意并回应他人需要的自然结果。但是“多样性、公平和包容倡议”（Diversity, Equity, and Inclusion）不是和谐伦理的特点——这些东西在陶然自得文化中是多余的。这样一个社区所关注的不是大原则，而是社区成员之间关系的具体属性。

九、非伦理的最后附言

基于美德、义务或有益结果的伦理不太适合AI和分析技术这样的领域。我们对何谓“好”意见不一，我们不能预测后果会是如何，我们不能事后解决不良后果造成的问题。伦理，特别是行业伦理，一般是以社会契约、权利或义务等进行界定的，因此体现为规定或原则。但是这些没有考虑背景和具体情况、相互关联的大环境以及分析技术和AI自身是如何运行的等方面的问题。

相反，从女性主义哲学观看，伦理，包括AI伦理，关乎的是关系，即我们如何相处和互相照顾。在我们应用分析技术时有一点很重要，即分析技术总是想成为我们的替身，正如文化人类学学者迈克·韦施（Michael Wesch）的视频“机器即我们/在利用我们”（the Machine is US/ing US）以及在微软聊天机器人Tay演变为种族主义AI的例子都提示了这种风险的存在。从具体、实际的意义上讲，AI伦理建立在我们做什么和对彼此说什么的基础上，如同我们想弄清楚诸如“为什么某人觉得发这样的帖子是合适的”这样的问题时所诉诸的伦理。如果AI伦理出问题，这只是反映出社会秩序有问题（Belshaw, 2011）。正是这个方面出问题了，我们才研究注意义务，这种女性主义哲学观采用强调关系和具体环境的方法研究道德和决策、分析在实际中行之有效的道德和伦理关系。其目标不在于“权利”或“公

平”而是在于诸如同情心这样的东西，不在于严苛的原则而是一种关心和友善的态度或方法，不在于约束或控制我们想做错事的诱惑而是寻找做好事的方法。

最后，源自我们自己生活经历的伦理观把社区视为一个完整的系统，任何决定都不是由某一个人做出的。我们必须谨记我们是如何建立关系的，更重要的是，我们首先是如何学习遵守伦理的（以别于那些定义何谓伦理的规定）。在实践、学习和工作中以及社会上应该如何做？我们认为可以通过以下途径达成这个目标：培育伦理文化（而不是强调遵守规定）、鼓励多样性视角培育更广泛的社区意识、鼓励开放性和相互影响培育同理心和从别人的角度看问题的能力。所有这些都不是伦理原则，但却是建设伦理社会的途径。

参考文献

因篇幅限制本文参考文献详见以下网址：

https://www.downes.ca/files/docs/2023_10_23_-_Ethics_Analytics_and_the_Duty_of_Care_-_Published_English.pdf

作者简介

史蒂芬·道恩斯（Stephen Downes），加拿大国家研究委员会（National Research Council Canada）高级研究员，联通主义和慕课始创者之一。

译者简介

肖俊洪，教授，《远程开放教育SpringerBriefs系列丛书》（SpringerBriefs in Open and Distance Education）主编，《开放、远程和数字教育期刊》（Journal of Open, Distance, and Digital Education）联合创刊主编。

Ethics, Analytics, and the Duty of Care

Stephen Downes

(National Research Council Canada, Ottawa, Ontario K1A 0R6, Canada)

Abstract: Artificial Intelligence (AI) and Learning Analytics have raised a host of ethical issues and a renewed attention to matters such as fairness, justice and benevolence. This article offers a comprehensive analysis of these topics, surveying the applications of AI and Analytics, with a focus on learning technology, and listing the ethical issues that have been raised. This is followed by an analysis of the ethical decision points that arise in the design of AI and Analytics, a study of relevant ethical codes for related professions, and an overview of the theories of ethics underlying those codes, leading to a contemporary analysis based in a philosophy of care ethics, and concluding with a discussion of ethical practices.

Keywords: analytics; AI; ethics; care; community; culture