Original Research Article

机器学习如何“思考”：理解机器学习的不透明性

Jenna Burrell
詹娜·布瑞尔*

内容提要 本文认为不透明性是一个涉及分类和排名的社会影响机制问题，如垃圾邮件过滤器，信用卡欺诈检测，搜索引擎，新闻动态，市场细分和广告，保险或贷款资格，信用卡评分。这些分类机制都经常依赖于计算机算法，并且在许多情况下依赖于机器学习算法来完成这项工作。在本文中，我对算法不透明性三种形式的进行了区分：（1）因国家机密与商业秘密而产生的不透明性，（2）技术无知产生的不透明性，（3）基于机器学习算法的特征和范围的有效应用产生的不透明性。本文深入分析算法本身。我引用了现有的计算机科学文献，行业惯例（已公开），并且以轻量级代码审计的形式对代码进行了一些测试操作。我认为，识别特定应用程序中不透明的形式，是确定各种技术和非技术解决方案中，防止损害的关键所在。

关键词 不透明性 机器学习 分类 不平等 歧视 垃圾邮件过滤

* 加州大学伯克利分校信息学院副教授。主要关注：边缘化社区如何适应技术、算法的公平性和不透明性、人类对算法的控制、民族志。研究领域：网络交际、社会与文化研究、发展中地区的技术、用户体验研究。

本文认为不透明性是一个涉及分类和排名的社会影响机制问题，如垃圾邮件过滤器，信用卡欺诈检测，搜索引擎，新闻动态，市场细分和广告，保险或贷款资格，信用卡评分。这些分类机制都经常依赖于计算机算法，并且在许多情况下依赖于机器学习算法来完成这项工作。

不透明性似乎是法学学者和社会科学家对“算法”的新关注焦点。所讨论的算法是对数据起作用的。使用这些数据作为输入，算法会产生一个输出，特别是产生一个分类（即是否向申请人提供贷款，或是否将电子邮件标记为垃圾邮件）。对于算法输入结果的接收者来说，算法（分类决策）是不透明的，很少有人会理解特定的分类是怎样或为什么从输入数据产生出来。另外，输入的数据可能是完全未知的或仅是部分已知的。问题自然而然地出现了，是什么样的原因导致了这种不清楚的状态？是因为算法的专有性吗？还是因为算法的复杂或高度的技术性？或其他原因？

通过厘清新兴的跨学科学术研究中含混不清的不透明性的不同形式，我尝试突出算法分类的不同内涵，回应社会学家长期关注的经济不平等和社会流动性等问题。三种不透明性的形式包括：（1）公司或机构有意而为的自我保护和隐藏，与之相伴的欺骗的可知性；（2）因写（或读）代码是一项专业技能而产生的不透明性；（3）机器学习高维度特征的数学优化与人类尺度的推理和语义解释之间的不匹配导致的不透明。不透明性的第三种形式（经常与第二种形式合并，从一般意义上讲，算法和代码具有高技术性和复杂性部分）是本文的重点。通过深入研究这种形式的不透明性，本文指出某些提议中将代码或算法“审计”作为评估歧视性分类的方法存在的缺陷。

为了研究不透明性的问题，深入理解算法本身，我引用了计算机科学中现有的文献、行业惯例（已公开），并以轻量级审计的形式对代码进行了测试操作。在此过程中，我将不透明性的形式与为解决机器学习分类的不可渗透性而提出的技术和非技术解决方案相关联。每种形式都提出了防止伤害的不同解决方案。

那么，有什么新东西？

近来，“算法”一词在公众场合的表达方式发生了转变，从一个几乎只在计算机科学家中使用的晦涩技术术语，变成了一个具有两极分化趋势的术语。其越来越多地出现在主流媒体上。例如，例如，美国护士联合会制作了一个广播节目（作者在当地的广播电台中听到），开头戏称，“算法是无人理解的简单数学公式”，结尾一位护士从一个对病人病情作出一系列滑稽错误声明的疾病诊断系统中跳出来拯救一个痛苦的病人[①]。这则公益广告（PSA）的目的是倡导（护士的）专业护理，在此情境下，反对容易出错的自动化。相比之下，企业为“算法”一词所做的“品牌化”努力，与偏颇的人类决策相较，强调算法的客观性（Sandvig，2015）。通过这种方式，该术语的含义被积极的塑造成广告文化和企业自我展示的一部分，同时也受到了与自动化、企业责任和媒体垄断相关的反话语的挑战（即Tufekei，2014）

尽管这些新媒体的叙述可能很新颖，但长期以来，大型组织（包括私营部门

① 参见 https://soundcloud.com/national-nurses-united/radio-adalgorithms。

公司和公共机构）采用的内部程序，受其约束者并不完全理解。这些程序可以被称为“算法”。那么，对于新术语用法以及随之而来的批评和分析，我们应该如何应对？这仅仅是“新瓶装旧酒”，还是伴随现实应用中越来越多地使用与算法相关设计模式，而产生的紧迫新问题？

除了关于算法的两极化公共讨论外，这一领域新出现的更多是普遍的理论技术和数据收集的应用技术，更庞大的包括与购买活动、链接点击和地理空间移动相关的，更普遍使用移动设备、服务、应用以及（世界局部）恒常连接的现状产生的个人数据档案。但是这与作用于数据的算法不一定关系密切。这通常关乎数据的组成，以及隐私和退出的可能性（或令人担忧的退出不能）的新关注。

其他的变化与特定的应用领域和不断演变的监管回应方案相关。将算法自动化转移到以前白领工作的新领域，反映在诸如“我们需要教师还是算法？”的标题中[②]，算法自动化也进入先前人为决定的相应分类过程，例如为了节省成本而进行的信用评估（常促使其向自动化领域转变）（Straka，2000）。Fourcade 和 Healy 指出，在信贷和借贷领域，发生了从以前的排他性借贷到可进行少数选择，再到更慷慨地提供借贷给更广泛的社会范围，但提供了一些不利的甚至高利贷的条款的转变。“追踪和分类消费者行为的方法的出现和扩展”使这些转变成为可能（Fourcade & Healy，2013：560）。这些方法（部分）是通过计算机中的算法得以应用的。这里的解释似乎表明，特定的算法程序拿下的工作领域正在扩大，算法正在以前所未有的规模承担更广泛类型的任务。

在法律学者和社会科学界对“算法”的这一新兴事物进行批判时，很少有人深入地考虑过其数学设计。反而采取了广泛的社会技术方法来研究“野外算法”。所研究的算法是基于他们在利润和股东价值压力下在公司内部的定位，以及它们被应用于特定的真实用户群体（以及这些用户产生的数据）的情形。因此更多超过算法逻辑的事物正在被检验。这样的分析通常特定于具有特定用户群的实现（例如 Google 的搜索引擎），并且独特地累积了问题和失败的历史记录，并提供了最终的参数设置和程序员的手动调整。这种方法可能不会在特定的算法类别中发现重要的泛化模式或风险。

不透明性探讨：一种方法和途径

一般来说，对于许多广泛使用的重要分类算法，我们不能直接查看代码。这种不透明性（在某种程度上来说）的存在是出于专有的考虑。它们封闭是为了保持竞争优势/保持领先于对手。对手可能是市场上的其他公司或恶意攻击者（与许多网络安全应用程序相关）。然而，可以通过教学资料了解这些算法使用的一般计算设计。

为了做到这一点，我从本科教育中使用的特别的机器学习模型部分提取了经典例证。在这种情况下，我专门研究了 Coursera 机器学习课程的编程任务。这些示例提供了简化的计算设想版本，可按比例缩小以在学生的个人计算机上运行，从而使其立即传回输出。这样的例子并不会带来许多棘手的现实应用程序挑战。换言之，不透明的方式尽管如此简化，但这种简化揭示了克服它的局限性的一些重要而根本的内容。

机器学习算法并未涵盖学者们感兴趣且正在研究的可能被置于“算法政治”

② 参见 Khosla (2012)。

[③]旗帜下的所有算法内容。然而，它们通常用于分类，被用于具有社会后果意义的预测，例如“此贷款申请人违约的可能性有多大？”因此值得特别考虑。在广泛的算法应用领域（如搜索引擎或信用评分），机器学习可能发挥中心或外围的作用，并不是总是容易区分哪一个是真实情况。例如，一个搜索引擎的请求是由算法驱动的[④]，但是搜索引擎算法并不是它们的核心“机器学习”算法。搜索引擎使用机器学习算法来达到特定的目的，比如检测广告或公然操纵搜索排名，以及根据用户的位置对搜索结果进行优先排序[⑤]。

虽然并非机器学习应用于所有任务都是分类任务，但这是一个关键的应用领域，也是一个引起许多社会学关注的领域。就像 Bowker 和 Star 写下的它们的分类描述和结果，“每一个类别限定了一些观点并压制了另一种”，生命遇上分类系统时被“破碎、扭曲、弯曲”，就像实行种族隔离制度的南非的种族分类系统，以及肺结核患者的分类（BowkerandStar，1999）。所谓算法会更‘客观’地进行分类（从而解决以往分类中存在的不足或不公正）的说法，不能简单地从表面上看，因为在设计算法时，人类的判断程度依然存在，选择会内置。这些人工工作包括定义特征、对训练数据进行预分类以及调整阈值和参数。

不透明性

在下文中，我定义了不透明性的类型，首先以“不透明性”作为一种专有保护形式或“公司秘密”（Pasquale，2015）。其次，在代码的可读性方面的不透明性。代码编写对于算法的计算实现是必不可少的技能，也是一项尚未被大众中广泛了解的专业技能。最后，为说明本文主旨，我对比了不透明性的第三种形式，其重点是机器学习算法的数学过程与人类语义解释风格之间的不匹配。这一挑战的核心是与机器学习中使用的特定技术相关的不透明性。每一种不透明的形式都可以用立法、组织、程序、技术等不同方式加以处理。但重要的是，必须识别特定算法应用程序中所包含的不透明性形式，以便采取行动来减轻问题。

不透明性的形式

公司或国家机密的不透明性

新出现的文献中有一种关于“算法政治”的观点认为，算法不透明在很大程

③ 这一领域的大多数学者关注于特定的应用领域，而没有说明所使用的算法的技术类别。Gillespie 着眼于搜索、趋势研究和其他包含过滤和排名算法的内容(2012 年)，Pasquale 着眼于声誉、搜索和金融算法(2015 年)，Brunton 研究了垃圾邮件过滤(2013 年)，而 Diakopoulos s(2013 年)的考量范围虽广，但均与数据新闻有关。Sandvig 关注搜索的同时简要考虑了计算机科学入门课程(2015 年)中教授的基本排序算法。Solon Barocas 专注的研究这一趋势的特别例外，主要着眼于机器学习算法的研究(Barocas, 2014a;Barocas, 2014 b;Barocas and Selbst, 2016)。

④ 除了部分内容（通常对用户完全不可见）之外，其他部分内容审核，交叉核对，实况调查和校正，可由人工进行。http://www.wired.com/2014/12/google-mapsground-truth/。

⑤ 在 Reddit AMA 上 Andrew Ng 谈到为什么公司要公开他们的算法技术的相关问题和解答参见 (https://www.reddit.com/r/Machine Learning/comments/32ihpe/ama_andrew_ng_and_ adam_coates/cqbkmyb)，有关机器学习如何为 Google 搜索做出贡献的问答和回应参见 http://www.quora.com/Why-ismachine-learning-used-heavily-for-Googles-ad-rankingand-less-for-their-search-ranking。

度上是企业为了维护自己的商业机密和竞争优势而有意为之的一种自我保护形式。然而，这不仅仅是一个搜索引擎和另一个竞争对手竞争来保持他们的“秘方”。也是占主导地位的平台和应用程序，尤其是那些使用算法进行排名、推荐、趋势分析和过滤的平台和应用程序，吸引那些想要将“博弈”算法作为吸引公众注意力的策略之一的人的平台和应用程序。“搜索引擎优化”领域就是这样做的。机器学习中有一种叫做“对抗性学习”的方法专门处理这类进化策略。机器学习的网络安全应用程序明确的处理垃圾邮件、骗局和欺诈，并保持不透明性，以确保有效性。Sandvig 指出，这种“猫捉老鼠的游戏”使得大多数算法完全不可能（或必须）向公众公开（Sandvig 等人，2014：9）。也就是说，开源软件是专有和封闭算法的一个明显替代。成功的商业模式已经从开源运动中出现。甚至在“对抗式学习”中也有可选项，比如针对 Apache 的 SpamAssassin 垃圾邮件过滤器。

另一方面，Pasquale 的分析更多持怀疑态度，他提出目前算法不透明的程度在许多应用领域可能不合理，是监管松懈或滞后的产物。在他的著作《黑箱社会:控制金钱和信息的秘密算法》（The Black Box Society：The Secret Algorithms that Control Money and Information）中，他认为，一种对抗的局面确实在发挥作用，其中的对手是监管本身。他说，‘如果金融家刻意让自己的行为不透明，就是为了规避或扰乱监管，那该怎么办？’他问道（Pasquale，2015：2）。关于这一点，他将“不透明性”定义为“可救济的不可知性”

Pasquale 认为，算法的不透明性可以归因于企业以追求竞争优势的名义故意而为的自我保护，但同时也可能是掩盖监管缺失、操纵消费者和/或歧视模式的新形式。

对于这种形式的不透明性，有一种应对建议是，必要时通过监管手段，使代码可供审查（Diakopoulos，2013；Gandy，2010；Pasquale，2015）。这种对算法不透明性的特殊解释是基于这样一个假设：如果企业愿意公开它们使用的算法的设计，那么通过阅读代码就有可能确定消费者操纵或违反监管的问题。Pasquale 承认，这些措施可能会使算法无效，尽管建议仍然可以使用独立的“可信的审计人员”，他们可以在服务于公众利益的同时维持保密性（Pasquale，2015：141）。在无法访问代码的情况下，Sandvig 等人（2014）详述并比较了几种形式的算法审计（无论公司是否合作的情况下进行）作为可能的应对，这是一种不需要访问代码本身就能强制解决问题的方式。

技术无知的不透明性

第二层次的不透明性源于承认目前编写（和阅读）代码和设计算法是一种非大众化的专业技能。软件工程课程强调编写简洁、优雅和易于理解的代码。虽然代码是用特定的编程语言（如 C 语言或 Python）实现，并且必须学习这些语言的语法，但它们在某些方面与人类语言大相径庭，其要求严格遵守逻辑规则，并保证拼写和语法的精确性，以便被机器“阅读”。

好的代码具有双重功能。它必须可以被人类（原始程序员、添加或维护代码的人）和计算设备（Mateas 和 Montfort，2005）解释。为计算设备编写程序需要特别的精确性、正式性和完整性，这非人类语言进行通信所必需。编程的技巧⑥是管理这种媒介作用，并需要一些众所周知的“最佳实践”，例如选择合理的变量名，包括“注释”（为机器编码时，省略了与人类程序员的单方通信），以及在

⑥ 参见 Ensmenger(2003)关于编程是一门技术和程序员是一种职业的观点。

所有条件都相同的情况下，选择更简单的代码格式。

最近呼吁在理工科领域增加多样性，并为在各级教育中努力发展“计算思维”（Lee 等人，2011；Wing，2006）。Diakopoulos（2013）同样提出，建议记者在逆向工程算法中发挥重要作用，向公众提供信息，但也指出，这对“人力资源”开发提出了挑战，即在记者或其他希望进行这类检查的人中开发代码和计算能力。为了解决这种形式的不透明性，深化教育将使公众更加了解这些可能影响他们生活机遇的机制，并更有利于他们直接评估和批评算法。

算法在应用范围内的操作不透明性

学者们已经注意到算法（比如在谷歌搜索引擎下的算法）通常是由团队构建的多组件系统，产生一种作为算法“内部人员”的程序员也必须应对的不透明性（Sandvig 等人，2014；Seaver,2014）。呼吁代码“审计”（即阅读代码）和聘用“审计员”可能低估了在一个复杂的软件系统中理清代码逻辑所需的时间代价。然而，这一有效的批判对于不同类型的算法及其特定的逻辑是泛化的。

我进一步指出，机器学习算法在规模和复杂性方面存在特有挑战。这些挑战不仅仅与代码的行数或页数、工程团队中的团队成员的数量，还与模块或子例程之间的大量互连有关。这不仅是阅读和理解代码的挑战，而且是理解操作数据的算法的挑战。尽管可以简单地以一种几乎可以完全理解其逻辑的方式来实现机器学习算法，但在实践中，这种情况不太可能特别有用。被证明有用的机器学习模型（特别是在分类的“准确性”方面）在一定程度上，不可避免的具有复杂性。

特别是机器学习经常被描述为遭受“维度诅咒”（Domingos，2012）。在“大数据”时代，数十亿或数万亿的数据示例和成千上万的数据属性（在机器学习中称为“特征”）可以被分析。当它“学习”训练数据时，算法的内部决策逻辑被改变。处理大量的数据，特别是异构的数据属性（即不仅仅是垃圾邮件中的文字，还有邮件头信息）会增加代码的复杂性。机器学习技术会随着规模的增长而迅速面临计算资源的限制，并且可以利用写入代码中的技术（例如“主要成分分析”）来管理这些资源，从而增加了其不透明性。尽管数据集可能非常庞大，但容易理解，并且可以清晰地编写代码，但两者之间在算法机制中的相互作用是产生复杂性（并因此带来不透明性）的原因。更好地理解这种复杂性（以及克服其影响的不透明性的障碍）是以下示例的关注点。

机器学习：一个非常简短的入门

机器学习算法被用作强大的泛化和预测器。由于我们知道这些算法的准确性会随着需要训练的数据量的增加而提高，因此近年来这些数据的可用性不断增长，重新引起了人们对这些算法的兴趣。

一个给定的机器学习算法通常包括两个并行操作，或两个不同的算法:一个“分类器”和一个“学习器”（例如，见图 3）。分类器接受输入（称为一组“特征”集）并产生输出（称为“类别”）。例如，进行垃圾邮件过滤的分类器采用一组特征（如电子邮件头信息、电子邮件正文中的单词等），并产生两个输出类别中的一个（“垃圾邮件”或“非垃圾邮件”）。进行疾病诊断的决策支持系统可以接受输入（临床表现/症状、血液测试结果），并将疾病诊断作

为输出（“高血压”“心脏病”“肝癌”）。然而，被称为“学习者”的机器学习算法必须首先对测试数据进行训练⑦。这个训练的结果是一个权重矩阵，分类器将使用它来确定新的输入数据的分类。例如，这些训练数据可能是预先分类并贴上“垃圾邮件”或“非垃圾邮件”标签的电子邮件。

机器学习包含许多模型，它们以不同的方式在代码中实现。一些流行的机器学习模型包括神经网络、决策树、朴素贝叶斯,和逻辑回归。模型的选择取决于适用的领域（即贷款默认预测和图像识别），它在其他方面证明了分类的准确性以及可用的计算资源。模型也可能被组合成“模型组合”，这是一种在机器学习竞赛中经常使用的方法，旨在分类准确性的最大化。下面将讨论使用不同模型的机器学习的两个应用。

在神经网络中可视化不透明度

我想讨论的机器学习的第一个模型和应用是一个应用于图像识别任务的“神经网络”。因为这是一项图像识别任务，所以可以尝试“看到”训练算法输出的权重。笔迹识别是向计算机科学专业本科生教授神经网络的经典例子。为了教学目的，简化计算任务，代码被实现为只识别手写数字（数字 0 到 9）⑧。为了进一步简化任务，这些数字被绘制在一个空间受限方框的边界内。查看图 1，可以看到要分类的数据的一些“模糊性”和不确定性。如果你把一个手写的数字写成 8*8 个像素的正方形，每个像素（以及与其相关联的灰度值）成为分类器的输入（或“特征”），分类器最终输出它所识别的数字（在图 2 中，它应该是数字 6）。

在神经网络的设计中，一组输入节点连接到第二组称为“隐含层”的节点（类似于大脑中相互连接的神经元），然后连接到输出层（见图 3）。每个输入节点连接到一个隐含层节点，在图 3 的神经网络设计中，隐含层的每个节点都连接到一个输出。一个值或权值与每一条连接的线相关联。

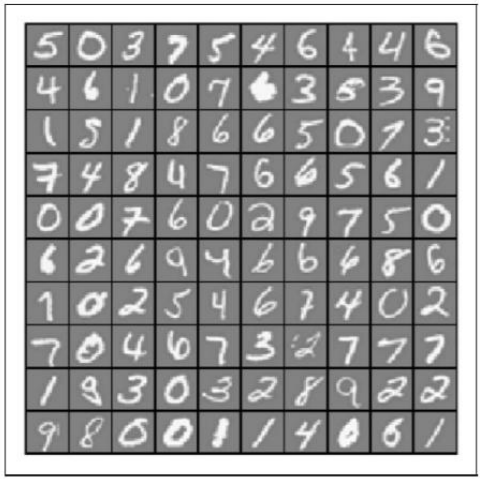

图 1.机器学习算法（“学习者”）一组手写数字的例子，神经网络可以在这种情况下得到训练。

⑦ 这里为了论证清楚起见，特别指出，此处是指被称为“监督学习”的机器学习方法的子集。

⑧ 我在 2001 年读本科的时候和 2013 年我完成的 Coursera 课程中用的是完全相同的例子，这或许能够更直观的感受到算法本身的变化有多么微小。

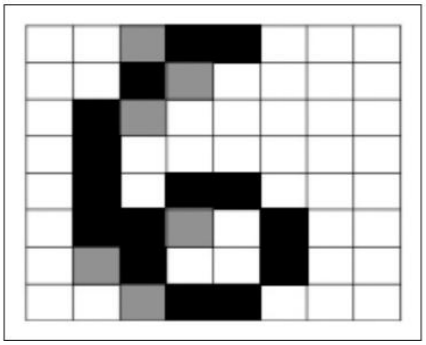

图 2.在 8×8 像素的正方形中的一个手写数字。

在图 3 中的神经网络设计中，隐藏层中的每个节点都连接到一个输出。值或权重与这些连接线中的每一条相关联。权值矩阵的最优值是学习算法学习的。“最优”是由一组权值来定义的，这些权值可以产生最精确的输入分类（在 8×8 矩阵中，单个像素及其强度范围从白到黑）到输出（这些像素代表的手写数字）。

因为这是一个图像识别任务，我们实际上可以将优化后的权重可视化到隐藏层节点中。通过这种方式，我们可以看到神经网络如何分解识别手写数字的问题（见图 4）。

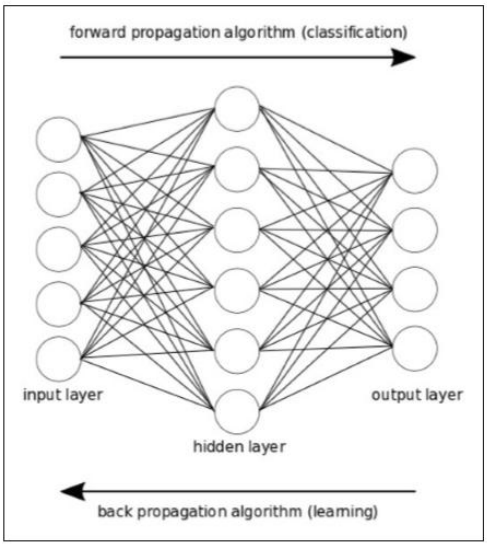

图 3.神经网络的图形化描述

图 4（a）说明了神经网络中的隐藏层。如果你看 25 个方框中的一个，你可以看到它提示了手写数字的哪个部分。每个框表示隐藏层中的一个节点，并且框中的每个像素表示从一个输入层节点到该特定隐藏层节点的权重值。总之，每个方框显示了一个只有一个隐藏层的简化神经网络的权重集。框中的黑色区域是相关节点最敏感的特定像素。例如，左上角的框显示了一个隐藏的层节点，提示了该节点出现在象限的左下方和中间的一点暗像素。这些隐藏层节点的计算组合会将输入分类为 0 到 9。

值得注意的是，例如，神经网络不会将手写数字识别分解为人类容易理解的子任务，例如识别一个水平条、一个闭合的椭圆形、一条对角线等。这个结果，这些权重中明显的非模式化，源于计算“学习”的概念。机器学习被应用于解决编码决策函数逻辑功能非常差的一类问题。在他的 Coursera 机器学习课程中，

AndrewNg 把这个描述为“我们不能用‘手’来编程的[应用]领域”⑨，“手”意味着人类的手⑩。如上所示，（人类）编写代码的技巧是双向通信，一方面是对人类程序员同胞，另一方面是对计算机处理器。当一个算法进行“编程”（即最优地计算其权重）时，逻辑上来说就会得出这样的结论:对人类而言，具有可解性（这是编写代码艺术的一部分）不再是一个问题，至少对非人类的“程序员”来说不是。

⑨ “欢迎”视频及 Coursera 机器学习课程参见 https://www.coursera.org/learn/machinelearning/lecture/RKFpn/welcome。

⑩ 手工编程(在分类决策的背景下)也需要明确地勾勒出决策的逻辑，尤其是关涉将数据放入哪个类别。这种以符号型人工智能 (Olazaran, 1996)而闻名的理性主义方法曾一度占据主导地位。它有时被提及，带着一点怀旧之情，作为老式 AI (GOFAI) (Winograd, 2006)，它以一种严格形式化的方式限定了知识的符号表征。然而，这种方法未能实现其早期的目的，在许多任务上表现不佳，导致 AI 在兴趣和资金减少的情况下进入“寒冬”(Grudin, 2006)。

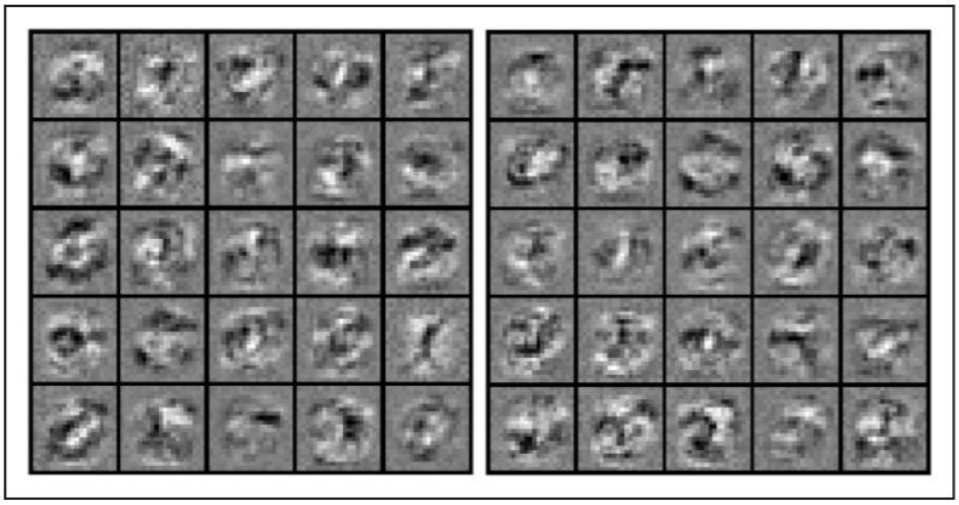

图 4（a）隐含层:每个框中的黑色区域是特定隐含层节点以手写数字表示的区域(笔画或其他图案)。（b）这是用相同的训练数据第二次运行相同的学习算法的结果。（a）和（b）不相同的原因是由于随机初始化步骤定义了一组初始值为非常小的随机数的权重。

第一个例子的主要目的是快速直观地了解机器的思维方式。图 4（a）应该显得不直观、随机、无组织。然而，手写识别对于人类来说也不是一项“有意识的”推理任务。人类对视觉元素的识别是以一种直接的、潜意识的方式进行的（因此在人类对文字的识别过程中也必然存在一种不透明性）。这个例子似乎并不能提供更多关于分类中歧视的更广泛的现实问题的见解。不过，最近的一个案例是谷歌照片中的自动分类。如果一组非裔美国人的照片被标为“大猩猩”⑪，那就说明事实并非如此。为了进一步说明这个论点，我的下一个例子，垃圾邮件过滤将介绍任务的自动化，它需要更有意识的人类推理形式。作为一个与互联网核心通信能力有关的问题，我展示了垃圾邮件过滤与分类歧视问题的相关性。

垃圾邮件过滤的不透明性

对垃圾邮件的定义不存在争议（Brunton，2013）。它通常被理解为不受欢迎的电子邮件，尤其是那些批量发送的电子邮件，但这在某种程度上，这是网络管理员对超负荷的网络资源的一种称呼。出于这个原因，垃圾邮件过滤是对于作为对社会产生影响的基于机器学习的分类来说是一个更好的应用领域。被归类为垃圾邮件的邮件信息是无法传递给其预期收件人的。因此，这个例子更直接地涉及到正在进行的关于搜索、排名和过滤内容的讨论。如果合法的邮件被归类为垃圾邮件（“假阳性”），那么实际上这封邮件已经在无意中被审查。一个问题是，垃圾邮件过滤器的设计是否能更容易将某些特定主体合法信息转移到垃圾邮件文件夹中。例如，若邮件来源于位于互联网欺诈或垃圾邮件活动的温床，西非（尼日利亚或加纳）或东欧，是否会导致该件被错误标记为垃圾邮件？

⑪ 事件详细描述参见 http://www.slate.com/blogs/future_tense/2015/06/30/google_s_image_recognition_ software_returns_some_surprisingly_racist_results.html。

在 Ng 的 Coursera 课程中，支持向量机（SVMs）是用于实现垃圾邮件过滤的机器学习模型。支持向量机是另一种类似神经网络的机器学习模型，任何一种模型都可以用于垃圾邮件过滤。Coursera 课程中使用的简化版本没有使用“内核技巧”（SVMs 的计算技术特征），因此它本质上是一种线性回归的形式；在技术术语中，它使用“线性核函数”。作为额外的简化，编程练习仅依赖于电子邮件的内容来训练垃圾邮件分类器，即只包含在邮件中的单词，而没有邮件头信息。“学习者”算法分析这些单词，以确定一组权重。这些权重衡量给定单词与“垃圾邮件”和“无垃圾邮件”（非垃圾邮件）电子邮件的关联程度。这样的方法被描述为一个“单词包”，单词之间没有假定的符号学关系，消息中没有任何意义被提取，也没有试图在算法中进行叙事分析。

我对算法提供了轻量级的“审计”，并对每个单词产生的权重进行了检查，以及我们可能如何理解它们。我特别关注一类垃圾邮件，即尼日利亚 419 骗局，这是一个我非常熟悉的类型（Burrell，2012）。419 骗局就网络访问和垃圾邮件的误报引发了一个有趣的关注。具体来说，地名，尤其是“尼日利亚”（Nigeria），是否会增加被归类为垃圾邮件的可能性？

事实上，在一个（公认的）非常过时的公共语料库⑫上运行训练算法后，可以生成一个地名列表及其相关的“权重”。这些权重在-1（与非垃圾邮件高度相关）到 1（与垃圾邮件高度相关）之间。令人欣慰的是，对于尼日利亚的电子邮件用户来说，电子邮件中包含的“尼日利亚”一词的权重是-0.001861。这意味着“尼日利亚”一词基本上是一个中性词⑬。纵观垃圾邮件的总体情况，这有一定的意义。总的来说，绝大多数的垃圾邮件都不是来自尼日利亚，也没有提到尼日利亚。据推测，提及尼日利亚的完全合法的电子邮件数量将进一步淡化该国与垃圾邮件之间的联系。

实际上与垃圾邮件最相关的单词（注意，这些单词已被去词尾，以便像 guarantee, guarantees, guaranteed 这些词可以作为等价词处理）如下：

our(0.500810)
click(0.464474)
remov(0.417698)
guarante(0.384834)
visit(0.369730)
basenumb⑭(0.345389)
dollar(0.323674)
price(0.268065)
will(0.264766)
most(0.261475)
pleas(0.259571)

在许多情况下，我们希望这些术语能够跨越垃圾邮件的类型。它们似乎暗示了一般

⑫ 自 2002 年以来 SpamAssassin 公开语料库参见 https://spamassassin.apache.org/publiccorpus/ readme.html。

⑬ 与垃圾邮件关联度从小到大的顺序排列，依次为爱尔兰(0.190707)、美国(0.108162)、华盛顿(0.076769)、波士顿(0.032227)、美国/美洲(0.015666)、印度(0.012690)、欧洲人/欧洲的(0.007351)、印度(0.006872)、欧洲(0.005295)、尼日利亚(0.001861)、法国(0.001398)、王国(0.027125)、外国(0.031424)、非洲(0.049945)、爱尔兰(0.062301)、加利福尼亚(0.067122)、单位(0.067960)、法郎(0.097339)、国家(0.101561)和中国(0.112738)。

⑭ 在对电子邮件内容进行预处理时，文本中的所有数字都替换为“基数”。

性的上诉、诉状和承诺（“保证”），集体的权威（“我们的”），以及具体和量化的收益或利益（尤其是货币）。

下面看一个具体的尼日利亚 419 风格的垃圾邮件示例，它最近被作者的 gmail 帐户垃圾邮件过滤器捕获，它确实被简化 SVM 垃圾邮件过滤器归类为垃圾邮件（完整的电子邮件见附录 1）：

我最亲爱的，

亲爱的，我是爱丽丝·沃尔顿夫人，美国公民。我给你带来了一个价值 1,000,000,000.00 美元的提议，我打算把它用于慈善事业，但我很害怕，因为在地球上很难找到一个值得信任的人……

在阅读这封邮件时，我注意到语言和单词的正式性，比如“最亲爱的”和“亲爱的”。提到“公民”、提供“慈善”、寻找“值得信任”的人，以及提到“欺诈”，也让人产生怀疑。然而，SVM 垃圾邮件过滤器不会提示这些单词。相反，在这封邮件中，提到金钱、“请”和“联系”是最重要的词语。事实上，在删除了邮件中提到的金钱和“请”这个词，并再次通过“分类器”算法进行处理后，它就不再被归类为垃圾邮件了。

现在对比一下，看看这封来自作者的朋友和研究合作伙伴的邮件，这封邮件有许多与诈骗邮件类型相同的标记(正式、虔诚、感谢的表达等等)，但不是诈骗邮件：

亲爱的教授，感谢您不断地使我恢复希望，并在所有希望都破灭的时候给我带来了生机。我带着泪水和深深的感激之情说谢谢……我能得到一个大发电机，空调，二手专业松下 3ccd 摄像机，还有 150 美元在我的帐户上可用于照顾我的健康……我祈祷你不断地成功。非常感谢，再见。

垃圾邮件分类器准确地将这封邮件归类为非垃圾邮件，同样完全基于它所包含的单词（不知道作者与发件人之间先前存在的关系）。尽管如此，当通过分类器算法运行时，电子邮件中会出现某些触发词（包括“想要”和“将要”），而且，最具罪证的是，还提到了钱。单词的“权重”排序似乎为解释人类思维的意义建构提供了一种杠杆，但即使在这个高度简化的例子中，也无法通过在特定电子邮件中简单浏览单词及其相关权重来确定整体分类。它是在电子邮件中找到的与 1899 年最常用单词的字典相匹配的所有单词权重的总和。细微的差异和关键词（即“访问”或“将要”）作为社会工程的垃圾邮件策略的一部分很难被理解，程式可能会打破分类的平衡。

人们可能会根据垃圾邮件的类型来识别和评估垃圾邮件：网络钓鱼骗局、尼日利亚 419 电子邮件、伟哥销售广告。相比之下，“单词包”方法将文本分解成单位的原子集合，这些单词的顺序与之是不相关的。该算法提供了垃圾邮件特有的非常普遍的术语，这些术语通常（单独的）非常普通和平庸。我的语义分析试图将算法呈现的统计模式与文本整体隐含策略相关的含义协调起来，但这显然不是机器“思考”的方式。

重新考虑“可解释性”

尼日利亚 419 风格的垃圾邮件分类的例子提供了一些关于代码审查的优点和缺点的见解。找到一些揭示算法内在逻辑的方法，可以解决人们对缺乏“公平性”和

歧视性影响的担忧，有时会有令人信服的证据证明算法的客观性，比如“尼日利亚”这个词的中性权重，进一步探究某一特定分类决策的“为什么”会产生一些暗示性的证据，这些证据似乎足以作为一种解释，但这将人类解释推理的过程强加给了统计优化的数学过程。换言之，机器思维被分解为人类思维的解释。然而，含糊不清的地方仍然存在，比如用“访问”和“想要”等无关痛痒的词作为垃圾邮件的指标。这让人怀疑，用这种方式来满足“为什么”问题的解释是否一定是一个特别正确的解释。

计算机科学家将这种不透明问题称为“可解释性”问题。“构建更多可解释分类器的一种方法是实现一个面向终端用户的组件，不仅提供分类结果，而且还公开这个分类的一些逻辑。”在垃圾邮件过滤领域，在谷歌 gmail 的“spam”文件夹中可以找到一个实际的实现。如果在此文件夹中选择了垃圾邮件，则会出现一个黄色警告框，提示“为什么此邮件属于垃圾邮件？”这封邮件的正文上方列出了它被放在这个文件夹的一个原因⑮。这些信息包括“它包含了垃圾邮件通常使用的内容”(可能指的是一种“文字包”的方式)和“许多人将类似的信息标记为钓鱼诈骗，所以这可能包含不安全的内容”。然而，一些解释提出了一份人类可管理的关键标准列表(比如，电子邮件中出现的 10 个权重最大的/垃圾词汇或单个句子描述)，提供了一种理解，往好了说是不完整的⑯，往坏了说是虚假的保证。

试图在“加权”输入和分类结果之间划一条直线的努力进一步复杂化的是，在这两者之间发生的数学操作。与这里介绍的手写识别和垃圾邮件过滤的例子不同，通常情况下，特征和模型中的维度之间的关系不是一对一的。操纵维数的方法(举两个例子，通过主成分分析或支持向量机中的“内核技巧”)经常被用来管理计算约束或提高精度。

计算能力的持续扩展产生了某些优化策略，随着规模的进一步复杂化，这些优化策略夸大了复杂规模的不透明性这个问题。随着更大的计算资源，以及需要挖掘的大量的 TB 级数据（现在通常是从用户活动的数字痕迹中随机收集的），分类器中可能包含的特征数量迅速增长，远远超出了推理人员能够轻易掌握的范围。在一篇关于应用机器学习的民间知识的文章中，Domingos（2012）指出“直觉在高维度下的失效。”换句话说，如果输入更多的质量或特征，那么推理、调试或改进算法将变得更加困难，每一种特征都会微妙地、不知不觉地改变结果分类。

有多种方法可以来处理这种基本的不透明性。可能令人惊讶的是，其中一种方法是避免在某些关键应用领域使用机器学习算法⑰。还有一些方法可以简化机器学习模型，例如“特征提取”，一种分析哪些特征对分类结果有实际影响的方法，从模型中删除所有其他特征。一些解决方案可能明智地放弃回答“为什么”问题，并

---

⑮ 关于这些解释的列表参见 https://support.google.com/mail/answer/1366858?hl¼en&expand¼5。

⑯ 一位计算机科学家同样提醒我们“有许多问题不能单靠简单、容易理解的决策理论来解决，这是我们使用机器学习而不是人工决策规则的根本原因。”Lipton(2015)。

⑰ 一位对开发自动驾驶汽车的研究人员，进行田野调查的社会科学家发现，这些研究人员完全避免使用机器学习，因为“你不知道它在学习什么”。无数训练集中没有出现的情况可能会导致不可预测，甚至危及生命的后果(Both, 2014)。在与作者的交谈中，雅虎公司和费埃哲公司(FICO 评分来源)也表示基于这个原因避免使用机器学习算法的情况。在信贷市场上，这体现的不仅仅是一种优先权，而是要通过《公平信用报告法案》来强制执行，该法案要求向消费者提供被拒信用的理由。然而，其他“替代性信用评分机构”或“消费者评分机构”可以自由地使用机器学习（ML）模型，而且（目前）不受这些规定的约束。详情请参阅 Cathy O’Neill 关于《数据科学实战》的教程 http://bclt.me/audio/Intro%20and%20 Keynote.mp3。

设计出可以以其他方式评估歧视的指标（如 Datta 等人，2015 年）。例如，在“通过感知实现公平”中，可以检测分类算法中的歧视性效应，而不必提取特定分类决策的“如何”和“为什么”（Dwork 等人，2011 年）。这在某种程度上扩展了 Sandvig 等人（2014）提出的外部审计方法和 Diakopoulos（2013）使用复杂的算法实现。

总结

算法最终可能会有一些难以理解的东西。（Gillespie,2012）

这篇文章的目的是更深入地研究机器学习算法及其“不透明性”的本质，并将其与分类和歧视中的社会利益联系起来。这是当下部分学术研究重新定位到“数字不平等”的问题，更多关注了计算资源和技能的分配（Hargittai，2008），但直到最近，也不是关注人们怎样受到计算分类，隐私入侵，或者其他在普通人群中不平等且可能违规的监视的影响（Barocas 和 Selbst，2016；Eubanks，2012；Fourcade 和 Healy，2013）。

对算法不透明性的法律批评通常集中在故意保密的能力上，并呼吁制定促进透明度的规则。Pasquale（2015）主张引入审计人员访问代码并确保分类的非歧视性。另一种方法是在更广泛的社会范围内培养代码编写和计算技能，以减少由同质的精英技术人员做出无法被非成员轻易评估的重大决策的问题。然而，机器学习算法的不透明性在更根本的层面上具有挑战性。当计算机学习并因此建立自己的分类决策表示时，它不考虑人类的理解能力。基于训练数据的机器优化并不理所当然地符合人类的语义解释。手写识别和垃圾邮件过滤的例子有助于说明，机器学习算法的工作方式如何能够逃脱人类的完全理解和解释，即使是那些受过专门训练的人，甚至是计算机科学家。

最终，法律学者、社会科学家、领域专家以及计算机科学家之间的合作，可能会解决因不透明性障碍产生的分类公平性这一具有挑战性的问题。此外，用户群体和公众可以对“领域专家”可能缺乏洞察力的排除和经验歧视形式（算法或其他形式的）发表意见[18]。减轻黑箱分类的问题不是依靠单一的工具或过程，而是依靠法规或审计（代码本身，更重要的是，算法的功能），使用更透明的替代方案（即开源），教育公众，以及提高有权编写重要代码的人的敏感性等组合方法。方法的具体组合将取决于给定的应用程序空间需要什么。

致谢

感谢在撰写本文的早期阶段审阅并提供意见的许多人，包括塞巴斯蒂安·本索尔、劳拉·德文多夫、什莱哈什·凯尔卡尔、马里昂·富尔卡德、迈克尔·卡尔·钱茨、索伦·巴罗卡斯、戴维·巴曼，史蒂夫·韦伯和加州大学伯克利分校社会科学矩阵“算法作为计算和文化”研讨会的成员。

⑱ 例如许多不同的群体在 Facebook 用以报告和核实的“实名制”政策机制实行中遇到了问题。https://www.eff.org/deeplinks/2015/02/facebooksname-policy-strikes-again-time-native-americans。

利益冲突声明

作者声明与本文的研究、作者身份和/或出版没有潜在的利益冲突。

基金

作者在研究、署名和/或发表本文方面没有得到任何资助。

参考文献

Barocas S (2014a) Data mining and the discourse on discrimination. In: Proceedings of the Data Ethics Workshop,Conference on Knowledge Discovery and Data Mining (KDD), 24–27 August, New York City.

Barocas S (2014b) Panic Inducing: Data Mining, Fairness, and Privacy, PhD Thesis, New York University, USA.

Barocas S and Selbst A (forthcoming) Big Data's disparate impact. California Law Review.

Both G (2014) What drives research in self-driving cars? (Part2: Surprisingly not machine learning). Available at: http://blog.castac.org/2014/04/what-drives-research-in-self-driving-cars-part-2-surprisingly-not-machine-learning/.Bowker GC and Star SL (1999) Sorting Things Out:Classification and Its Consequences. Cambridge, MA:The MIT Press.

Brunton F (2013) Spam. Cambridge, MA: The MIT Press.

Burrell J (2012) Invisible Users: Youth in the Internet Cafes of Urban Ghana. Cambridge, MA: The MIT Press.

Datta A, Tschantz MC and Datta A (2015) Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. In: Proceedings on Privacy Enhancing Technologies, 30 June–2 July, Philadelphia, PA.

Diakopoulos N (2013) AlgorithmicAccountability Reporting:On the Investigation of Black Boxes. Report, Tow Center for Digital Journalism, Columbia University.

Domingos P (2012) A few useful things to know about machine learning. Communications of the ACM 55(10): 78.

Dwork C, Hardt M, Pitassi T, et al. (2012) Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference,8–10 January, Cambridge, MA, pp. 214–226.

Ensmenger NL (2003) Letting the ''computer boys'' take over: Technology and the politics of organizational transformation.International Review of Social History 48 (S11):153–180.

Eubanks V (2012) Digital Dead End: Fighting for Social Justice in the Information Age. Cambridge, MA: The MIT Press.

Fourcade M and Healy K (2013) Accounting, organizations and society classification situations: Life-chances in the neoliberal era. Accounting, Organizations and Society 38(8): 559–572.

Gandy OH (2010) Engaging rational discrimination:Exploring reasons for placing r egulatory constraints on decision support systems. Ethics and Information Technolo gy 12(1): 29–42.
Gillespie T (2012) The relevance of algorithms. In: Gillespie T, Boczkowski P and Foot K (eds) Media Technologies:
Essays on Communication, Materiality, and Society. Cambridge, MA: The MIT Pre ss.
Grudin J (2006) Turing maturing: The separation of artificial intelligence and hum an–computer interaction. Interactions
13(5): 54–57.
Hargittai E (2008) The digital reproduction of inequality. In: Gursky D (ed.) Socia l
Stratification. Boulder, CO:Westview Press, pp. 936–944.
Khosla (2012) Will we need teachers or algorithms? In:TechCrunch. Available at: http://techcrunch.com/2012/01/15/teachers-or-algorithms/ (accessed 11 December 2015).Lee I, Martin F, Denner J, et al. (2011) Computational thinking for youth i n practice. ACM Inroads 2(1): 32–37.
Lipton Z (2015) The myth of model interpretability. Available at:http://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learn ing.html (accessed 11 December 2015).
Mateas M and Montfort N (2005) A box, darkly:Obfuscation, weird languages, an d code aesthetics. In:Proceedings of the 6th Annual Digital Arts and Culture Conference, 1–3 December, Copenhagen, Denmark..
Olazaran M (1996) A sociological study of the official history of the perceptrons controversy. Social Studies of Science 26(3): 611–659.
Pasquale F (2015) The Black Box Society: The Secret Algorithms that Control M oney and Information. Cambridge, MA: Harvard
University Press.
Sandvig C (2014) Seeing the sort: The aesthetic and industrial defense of ‘‘the algorithm’’. Journal of the New Media
Caucus 10(3): 1–21.
Sandvig C, Hamilton K, Karahalios K, et al. (2014) Auditingalgorithms: Research methods for detecting discrimination on internet platforms. In: Annual Meeting of the International Communication Association, Seattle, WA, pp. 1–23.
Seaver N (2014) Knowing algorithms.
Presented at Media in Transition 8, Cambridge, MA.
Straka JW (2000) A shift in the mortgage landscape: The 1990s move to automat ed credit evaluations. Journal of Housing
Research 11(2): 207–232.
Tufekci Z (2014) The year we get creeped out by the algorithms. Available at: htt p://www.niemanlab.org/2014/12/the-year-we-get-creeped-out-by-algorithms/ (accessed 17 June 2015).
Wing JM (2006) Computational thinking. Communications of the ACM 49(3): 33–35.
Winograd T (2006) Shifting viewpoints: Artificial intelligence and human–computer interaction. Artificial Intelligence 170(18): 1256–1258.

附录 1

作者 gmail 账户中垃圾邮件文件夹中的垃圾邮件：
我最亲爱的：

问候你，我亲爱的，我是爱丽丝·沃尔顿夫人，美国公民。我给你带来了一份价值 1000,000,000.00 美元的提案，我打算把它用于慈善事业，但我很害怕，因为在地球上很难找到一个值得信赖的人。我很高兴认识你，但是上帝更了解你，他知道为什么他在这个时候把我引向你，所以不要害怕。我知道有很多欺诈行为发送这样或其他形式的信息。我在商务部和外贸部看到了你的电子邮件联系方式。

我怀着沉重的悲痛写这封信给你，让你知道我患心脏病已经 22 年了，就在几个星期前，我的医生告诉我，我活不了多久了。

我正在联系你，因为我很感动地向你敞开我的项目。如果您有兴趣请回复我，如果不感兴趣请忽略此消息。

上帝保佑你。

如果您感兴趣，请回复我，以便我可以为您提供进一步的细节。

电子邮箱：alice.walton2@yandex.com

加纳研究员和朋友发来的非垃圾邮件：

亲爱的教授，感谢您不断地恢复希望，并在所有希望似乎都破灭时给我带来了生机。我带着泪水和深深的感激之情对你说声谢谢。我迟迟没有回复，因为我不想告诉你我还没有好转。我对治疗有反应，尽管不是很好。我能得到一个大发电机，空调，二手专业松下 3ccd 摄像机，还有我的帐户里照顾我的健康的 150 美元。我还把我的手机从一个有问题的旧诺基亚换成了一个 h6techno（先进的中国手机）。医生说我有疟原虫。只有上帝知道我什么时候才会好起来。我无法想象没有你生活会怎样。我祈祷你不断地成功。非常感谢，再见。

Abstract: This article considers the issue of opacity as a problem for socially consequential mechanisms of classification and ranking, such as spam filters, credit card fraud detection, search engines, news trends, market segmentation and advertising, insurance or loan qualification, and credit scoring. These mechanisms of classification all frequently rely on computational algorithms, and in many cases on machine learning algorithms to do this work. In this article, I draw a distinction between three forms of opacity: (1) opacity as intentional corporate or state secrecy, (2) opacity as technical illiteracy, and (3) an opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully. The analysis in this article gets inside the algorithms themselves. I cite existing literatures in computer science, known industry practices (as they are publicly presented), and do some testing and manipulation of code as a form of lightweight code audit. I argue that recognizing the distinct forms of

opacity that may be coming into play in a given application is a key to determining which of a variety of technical and non-technical solutions could help to prevent harm.

翻译：李龙帧 朱枫宇
校对：党家玉 孟宇辰