

Ethics of Big Data

大数据伦理

平衡风险与创新

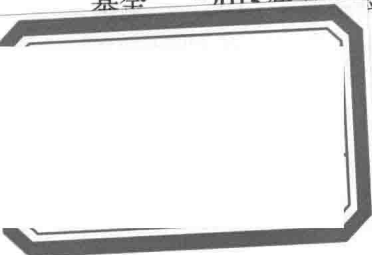


O'REILLY®

東北大學出版社
Northeastern University Press

[美] 科德·戴维斯 道格·帕特森 著
赵亮 王健 译

该译著受“辽宁省科技项目教育厅一般项目”(项目批号W2015147)、“辽宁省科技项目社科项目省社科规划项目”(项目批号L15BZX006)、“陈昌曙技术哲学发展基金”“2015年基本科研业务费追加项目”(项目批号L1513004)资助



大数据伦理

——平衡风险与创新

科德·戴维斯 道格·帕特森 著

赵亮 王健 译

东北大学出版社

·沈阳·

© 2016

图书在版编目 (CIP) 数据

大数据伦理：平衡风险与创新 / (美) 科德·戴维斯, (美) 道格·帕特森著；赵亮, 王健译. —沈阳：东北大学出版社, 2016.6

ISBN 978-7-5517-1313-9

I. ①大… II. ①科… ②道… III. ①数据处理—伦理学 IV. ①TP274

中国版本图书馆 CIP 数据核字 (2016) 第 144364 号

出 版 者：东北大学出版社

地址：沈阳市和平区文化路三号巷 11 号

邮编：110819

电话：024-83680267 (社务部) 83687331 (市场部)

传真：024-83687332 (总编室) 83680178 (出版部)

网址：<http://www.neupress.com>

E-mail: neuph@neupress.com

印 刷 者：沈阳中科印刷有限责任公司

发 行 者：东北大学出版社

幅面尺寸：178mm×233mm

印 张：4.5

字 数：96 千字

出版时间：2016 年 6 月第 1 版

印刷时间：2016 年 6 月第 1 次印刷

组稿编辑：刘振军

责任编辑：孙德海

责任校对：潘佳宁

封面设计：张 健

责任出版：唐敏志

ISBN 978-7-5517-1313-9

定 价：25.00 元

O'Reilly Media, Inc.介绍

O'Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O'Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly 的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly 为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 Make 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly 的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，在线服务或者面授课程，每一项 O'Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar 博客有口皆碑。”

——Wired

“O'Reilly 凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O'Reilly Conference 是聚集关键思想领袖的绝对典范。”

——CRN

“一本 O'Reilly 的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim 是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照 Yogi Berra 的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去 Tim 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

序言

哲学与商业并不总是协调发展。哲学一般不很关注调查研究蕴藏的实践意义，相反，商业则常常只对其经营策略的实际结果深感兴趣。

“伦理”是一个过于沉重的字眼。即便作为一个正规的学术研究领域，伦理内涵的各种概念预设也常常使人羞于把它当作一个讨论的话题。谈论我们尚不完全理解的东西是困难的，况且这个字眼本身有时就已包含了判断——类似“要这样做”“不要那样做”的指令和义务要求。而当我们认为自己正在被强令时，常常会被激怒。

本书试图在一定程度上解决这些困难。倒不是因为这些困难本身难以克服（虽然伦理学方面的调研确实是辛苦的工作），而是因为这些困难会在帮助各类组织从哲学性的思索与探究中获益上制造藩篱，而这种获益本应是极其丰厚的。我的方法的主要特色在于，能洞察出商业环境、市场、公司、文化、地缘特点及组织规模与成熟度等一切方面所造成的一系列不利的、复杂多变的组织环境因素。这些从你自身情况出发并不陌生的组织环境因素，对你的组织如何结合伦理调研实现最优化运营具有理所当然的重要影响。

人们常问我：“你如何以哲学学习背景来从事咨询工作？”我的回答这些年来在不断变化和充实，这主要是我从如何回答这个问题当中学到越来越多的东西，并且问题的答案在宏观抽象层面上关乎哲学与商业之间的关系，在微观具体层面上则又关乎伦理与大数据之间的关系。

我对技术的兴趣始于上五年级时，那时祖母送给我一套七十五合一的电子配件玩具，它的老版本至今仍能在eBay上买到！它告诉我，那些电线、电池、电容、电阻都能被不断组装和重新组装，从而创造出各种新的电路，实现多种多样奇妙有趣的功能。经过中学编程课程的学习，及后来在美国海岸警卫队的近5年中大多数时间里作为一名无线电技术人员对电信技术的钻研，我逐渐意识到，技术的迷人之处在于它直击我们的物质世界当中某些重

大并且复杂的现象之本质。能量能被定向传输，无线电波产生并能传送数字信息，晶体管及其他半导体材料能被组装成强大的新式计算处理和存储设备，软件能使所有这些设备发挥神奇的作用。

你大概以为我是学习物理专业或计算机科学的。事实上是哲学通过展现形上世界同物质世界的美妙之处而俘获了我的注意力——分析探究一切事物本质的那种严谨性，及洞悉科学技术自身发展规律的开放自由的方法。我不由自主地被深深吸引。

写作此书的一个主要目的是，应用一般而言的哲学方法、具体而言的伦理探究方法，帮助我们发展技术，并使其成为让我们生活得更舒适、更美好的工具。

下面说说大数据。这方面技术的独一无二之处恰恰在于其独有的本质特征，就是海量数据信息的产生、连接、关联、聚合、存储、处理、发布等。作为人类，我们从未面对过如何去应付如此海量数据的问题，更何况去理解这个问题。特别是针对人类及其行为而言，积聚如此海量数据信息的意义之一便是我所谓大数据“强加性的巨大功能”。不管我们是否欣然接受，这种强加性的巨大功能推动我们去思考一些严肃的伦理问题，包括对大数据的某些特定利用方式是否侵犯了基本的文化权利、社会权利、政治权利和法定权利。

这些问题是长久的、复杂的、深入的重要议题。我们从社会角度出发对它们的关注还不够充分。但也难以充分，因为我们并不习惯于在商业语境下充分展开这些议题。还是那句话，伦理已是一个过于沉重的字眼。正因为如此，你可以寄希望于，这本书将会帮助你和你的组织着手开发在新的背景下以新的方式开展明晰的伦理调研的能力，这也是本书试图达到的目标。本书中的新方法、新理念、新词汇试图从头开始帮助你获得更高的能力，使你能在自身所处的具体环境中去决定如何最有效地利用和执行明晰的伦理调研取得的成果，以助益于你的组织的发展。

就洞悉那些伦理议题及其结论而言，上述这种探讨还很初步。我们都只是随情况边发展变化边开始着手应对，所处的环境让我们别无选择。之前从未有人曾拥有过今天大数据所赋予的如此巨大的创新机遇，也从未有人面临过今天大数据所带来的如此巨大的意外风险。我们都只能在一个别无选择的境遇中边前进边探索。

我希望成为这一正在不断进行中的探讨过程的一部分。奥莱利出版社建设了一条专业出版流程，使本书能够通过按需印刷和数字通道等方式更方便地再版及重印。在对大数据的共同认知和应用进程中，这些建设工作使本书能够不断随之翻新升级。

本书中的惯用格式

本书中使用如下格式约定：

斜体字——表示新出现的术语、URL、电子邮件地址、文件名与文件扩展名。

示例代码的使用

本书中提供示例代码的目的是帮助你更好地完成学习。一般情况下，你可以在你的程序及文档中使用本书中的代码，而不必取得我们的许可，除非你要复制其中很大一部分。例如，你在编写程序中用到了来自本书的几个代码片段，这不必取得我们的许可。而如果将 O'Reilly（奥莱利）图书中的代码制作成光盘及进行售卖，则务必取得我们的许可。引用本书中的内容及示例代码来解决问题无需得到许可。将本书中很大一部分示例代码整合进你的产品文档则务必取得我们的许可。

如果引用了本书内容并声明内容的版权归属，我们对此表示赞赏，不过这并非必要。版权归属声明通常包括标题、作者、出版社及 ISBN 号，例如：“*Ethics of Big Data* by Kord Davis with Doug Patterson (O'Reilly). Copyright 2012 Kord Davis, 978-1-449-31179-7”。

如果你认为你对示例代码的使用情况不在上述给出的正常使用或许可范围，请随时联系我们：permissions@oreilly.com。

联系我们

如对本书有任何建议和问题，请通过以下方式联系我们：

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街2号成铭大厦C座807室（100035）
奥莱利技术咨询（北京）有限公司

针对本书我们有一个专门网页，用于列出勘误表、示例及其他补充信息。你可以通过下列网址访问该网页：

<http://oreil.ly/ethics-big-data>

如希望评论或询问与本书相关的技术问题，请发送电子邮件到以下邮箱：

bookquestions@oreilly.com

如想了解关于O'Reilly的图书、课程、会议及新闻等的更多信息，请访问我们的网站：

<http://www.oreilly.com>

致谢

本书受益于大量、广泛的人思想、投入与努力。我想要向他们中的几位致谢，并提前向那些可能忽略掉的人致歉。

首先，感谢ExactTarget公司的Brian Smith, Rob Wiley和Tom Williams。ExactTarget是一家不但出色地致力于将其价值观融入到其组织文化中去，而且在积极学习如何将大数据技术转化为有用工具方面处于前沿位置的公司。在很多方面，与他们一同工作的经验促成了这本书的写作动机。在大量的谈话、采访、晚宴和午餐时间里迸发出了大量的想法和素材，希望我已将我们的讨论成果在本书中充分呈现出来。

如果不是因为这些经由与XPLANE这家视觉思维公司合作而给了他们一个平台的项目，那些讨论就不可能发生。在XPLANE有很多出色的普通人以及天才，其中有几个人对本书做出了直接的贡献，他们是Matt Adams, Stephanie Gioia, Dave King和James Macanufo。

除了那些与我一同紧密合作的人，还有多种领域的专家，他们来自各种各样的产业和学科，并热心地献出他们的时间和专长来帮助解决各种各样的概念及其意义问题。这些专家包括Andrew Davis, Nigel Ballard, Jason Bobe, Erin Conroy, Pete Forsyth, Ezra Gollongly, Dion Hinchliffe, Erik Huddleston, Bill Hoffman, Max Niederhofer, Martha Koenig和Adam Riggs。

需要特别提到一个提供了大量专业意见的人，他就是Doug Patterson。他在学术上宝贵的背景、素养与专长激发了本书中的大量哲学思考。他教授商业伦理及就高端概念主题推动课堂讨论的经验，使他能够快速抓住关键性的伦理议题。在那些我需要使让各种议题变得复杂的各种更加细微的方面清晰化的时刻，他是一笔可以求助的强大资源。

还有一个专业的、渊博的、严谨的技术评议组织，给予了这本著作所应受的锤炼，我也希望他们的意见与付出得到应有的回馈，因为我知道他们使这本著作更经得起考验。因此，我要特别感谢Terence Craig, Bob Gourley, Mary E. Ludloff, James Macanufo和Cathy O'Neill。

最后——当然，并非最不重要——我要感谢我周围的很多朋友和同伴。我要从心底里感谢他们给予我的鼓励、信任、交流、忍耐、支持、助益，以及他们对这本书的写作价值从未间断的肯定。他们是：Laura Allen, Jake Baker, Tad Bamford, Cari Carter, Collin Con-

non, Tanya Frantzen, Patrick Foss, Vincent Grace, Drew Hansen, Erica Hassinger, Geoff Rogers, Jodi Sweetman, Carisa Sprecher, Khrist Soden, Ben Thompson, Paul Wille, Rob Woolsey 和 Morgan Wu。

最后的最后，要把一份极大的感激和公开的谢意送给 O'Reilly 出版社的广大同人，他们的付出已成为这项工作成果的一部分。

特别是我的主编 Courtney Nash，当我告诉她我正在计划写一本个人出版的大数据伦理方面的书时，她立即着手调查是否有其他人已经与 O'Reilly 方面签署类似的写作计划，并帮助我向编辑部提交投标计划。特别要感谢她能注意到这一主题的价值，摘得头筹，并在保证著作质量的同时，在进行过程中为帮助确保这个项目持续高效的展开而勤勉工作。

如果没有临时编辑（此时 Courtney 正在进行一项她自己的私人项目——生出一个美丽的女婴）Julie Steele，这本著作也同样无法面世。Julie 在一个非常繁忙和关键的时段淡定地参与进来，并在工作继续推进中帮助确保我始终朝向正确的方向。

当然，最终还要感谢 Tim O'Reilly 先生创办了这样一家出版机构，它将出版工作聚焦于这样一类有意义的主题，并致力于以他的观点与境界关注技术、文化和社会。

希望大家喜欢这本书，希望这本书对大家有用。

目录

序言	1
第1章 大数据：大冲击	1
1.1 大数据为何“大”？	4
1.2 大数据“强加”了什么？	5
1.3 大数据是伦理中立的	8
1.4 不要告诉我该做什么	10
1.5 重要概念和术语	10
第2章 价值观与行动	12
2.1 清晰表述你的价值观	13
2.2 让价值观转化为行动	14
2.2.1 大数据伦理的四块基本构成：身份、隐私、所有权与名誉	15
2.2.2 伦理探究的益处	18
2.2.3 价值观与什么相关？	19
2.3 伦理决策点方法	21
第3章 目前的实践情况	26
3.1 研究结果概述	27
3.2 数据买卖情况对比	28
3.3 “选择性拒绝”“选择性接受”情况对比	29
3.4 通过聚合进行关联	30
3.5 数据所有权	33

3.6 价值观的彰显	34
3.7 伦理上的失调	35
3.8 其他名义的政策情况	35
3.9 文化层面上的价值观	38
3.10 我们的学习成果	38

第4章 统一价值观与行动39

4.1 方法和工具	39
4.2 统一价值观与行动的方法论框架	41
4.2.1 调查	42
4.2.2 分析	44
4.2.3 表述	50
4.2.4 行动	51
4.2.5 价值观的角色化	52
4.2.6 GDM公司：一项个案研究	55
4.2.7 价值观/行为统一的有益之处	58

关于作者60

大数据：大冲击

一个将特权置于道义之上的人会很快将二者都失去。

——德怀特·戴维·艾森豪威尔

我选择用我的作品作为我价值观的体现。

——西德尼·波蒂埃

美国塔吉特公司知道，苹果公司知道，领英 (LinkedIn)、奈飞 (Netflix)、脸谱 (Facebook)、推特 (Twitter)、Expedia等互联网巨头也知道，各大国家及地方政治团体，乃至数以百计的其他那些创造着巨大的经济、政治与社会价值的组织都知道：大数据时代来了，并将一发而不可收！越来越多知名或不知名的企业组织加入到越来越依赖大数据技术的组织行列中来。

2012年2月16日，《纽约时报》登载了一篇关于塔吉特公司能够识别出消费顾客是否怀孕的文章。塔吉特公司拒绝评论或出面解释这个故事，但它还是被写作完成并登载出来。随后，围绕从塔吉特公司商业行为的合法性问题，到公众对隐私、个人信息等公共问题更加广泛的公开关注，攻击性的评论和后续报道接踵而至，引发出大量质疑的声音。

2011年4月20日，两名安全调查员宣布，他们发现苹果iPhone手机已有计划地将每部设备的位置信息完整记录到了一个隐藏文件当中。苹果公司方面大方承认该情况属实，而与此同时，随之而来的骚动表明，是该文件被生成并存储的方法问题，引发了人们对安全问题的强烈关注。在线下的现实世界中使用这种技术手段的决策，会造成明确直接的伦理后果。谁参与了这一决策的做出？是关在密室里两耳不闻窗外事的技术工程师在扮演着最重

要的角色吗？就那些功能应否投入使用的问题，到底是否存在着更广泛的商业讨论？安全性及其他风险问题应该在什么具体层面上得到讨论？

2011年8月，当脸谱（Facebook）考虑通过其移动终端应用中的“联络”特色功能，在用户的移动通讯设备上公开所有用户的姓名及电话号码，以方便联络时，面临了舆论的批判。脸谱方面回应并澄清了该特色功能的实现方式，并向用户提供了从他们的脸谱账户中删去这些被公开个人信息的方法。但为什么这种澄清及方法提供不是同步于功能发布的第一时间呢？

2011年，据报道，当美国提供域名注册和互联网主机服务的GoDaddy公司的CEO鲍勃·帕森斯在推特上发表了一条自己在非洲猎杀大象的推文^{注1}，又公开支持了引起强烈争议的《禁止网络盗版法案》（SOPA）^{注2}后，引发了其用户的负面回应，导致其域名注册业务流失了数以万计的客户。当美国品牌购物网Kenneth Cole旗下品牌被发现利用2011年春天开罗发生的动乱来促销其产品后，立即招致毁灭性打击，紧随其后的道歉也无法挽回一个品牌声誉终结的下场。2010年，沃尔玛被指控利用一个虚假的在线社区来伪造支持的声音，以达到在接受度尚不充分的地区建立新连锁店的目的，一家公关公司被指控对此负责。

当你也许正在思考你的组织将如何应对类似情况时，请注意这样一个事实：以上全部例子都牵扯到同一个因素——大数据技术。正如这些例子所显示的，大数据的一个冲击便是行为会以更加迅捷的速度，产生更加巨大深远的后果，并直接影响到一个企业的品牌价值、客户关系和收益所得。因此，大数据正在迫使我们围绕价值观与行为活动展开新的思考，尤其是当大数据为更多的人提供更多的途径进行参与、交流和互动时。大数据技术的这种不断加剧的趋势的一个结果便是，当经营活动不断变革并持续快速生成海量数据信息的同时，大数据现象也正在开始引发伦理问题。

正如布拉德·彼得斯在《福布斯》杂志上所述，大数据“改变了社会契约形式”（<http://www.forbes.com/sites/bradpeters/2012/07/12/the-age-of-big-data/>）。这一变革的性质是复杂的。本书写作的一个主要动机便是同时聚焦个体和组织，提出更加明确与透明的、内在包含伦理考量的讨论之必要性。

注1： 相关事件的报道及部分评论见以下两个网址：<http://it.sohu.com/20110401/n280084849.shtml>
<http://www.chinaz.com/news/2011/0401/169348.shtml>（译者注）

注2： 其具体内容略见于以下两个网址：

<http://baike.baidu.com/link?url=dd0EA9joCOS3nMm3EQ2fXuCZMg6DdoYSMZm3DqQQ-TAt-apsbGhMms23JSQ7co2Oi3-ZAxSZBy7Jpa0omcRXZm1rrluAp-h2gJX6ILmaFaHu>

http://baike.baidu.com/link?url=I5An0OEdeMhN1sd_6zdeEYLhBD_baOv05r_LiKUdOg-gCUE5XJRfR8UyqW3qqrYmex5nCxxwYliUnzzbS4IUk_（译者注）

尽管一些伦理论题主要与个体的人有关，其隐含的涉猎范围仍跨越多种领域。同样，大数据引发了对个人隐私的忧虑，也造成了新的个体身份问题，特别是谁对我们的私人数据具有所有权，以及海量数据与日俱增的涌现与可利用性会对我们的声誉产生怎样的影响。

无论对于个体还是组织，以下4个基本要素都能用以定义出一个可被视为大数据伦理的框架：

身份：我们的线下身份与线上身份之间是什么关系？

隐私：谁应该来控制数据的使用权？

所有权：谁是数据的所有者？谁有权转让数据？产生数据和使用数据的人履行什么义务？

名誉：我们如何断定数据的信实性？无论是对于我们自己、其他人，还是任何别的事物，大数据以指数级增加着信息量和互动方式。这种情况使我们如何被认识与评价这一问题的复杂性大大增加。

无论个人还是组织都拥有知晓数据正如何被操作的合法权益。不管你在一个组织中扮演什么角色，哪怕你在从事技术工作，今天几乎所有人的生活都被大数据技术所触及。这就意味着上述框架既具有预示大数据所提供的利益的潜在作用，又具有预示源自意外后果的潜在风险的作用，而这意外后果将影响到绝大多数人。

例如，纽约法官加里·布朗发现，一个IP地址并不能作为指证版权侵害的充分证据(<http://torrentfreak.com/judge-an-ip-address-doesnt-identify-a-person-120503/>)。尽管该法律上的发现结果聚焦于版权问题，但它却能被推而广之到前面大数据伦理框架中全部4个元素所涉及的问题。如果一个IP地址无法对应到一个具体的人（谁曾经真的以为二者之间是一一对应的吗？），那么经由一个特定IP地址所产生的数据还能被正当合理地跟一个特定的独立个体联系在一起吗？

利用网络进行市场营销的销售人员对这种状况头痛已久。但自从牵涉到大数据后，意外后果的风险形成更加广泛的影响，已经完全超越目标营销领域。可以深入思考一下，如果你同一个或几个人共同按时间平均使用同一台家用电脑，那么Google该怎样过滤识别你对网络内容的偏好呢？笔者对海滩度假地点的喜好，就跟与笔者共享同一网络连接的人大为不同——此人惧怕大海且不会游泳。提高目标营销的准确性是一大挑战，而这一挑战及其潜在风险不会以线上广告作为终点。

一个现实场景可以说明人和组织所要面对的一些挑战。我们来想象一位上了年纪的亲属的葡萄糖与心脏追踪监测设备和你家庭内部其他人的监测设备共用同一个IP地址。自然而然，来自该种医疗设备的全部数据都被某个卫生保健服务供应方捕捉和存储。现在试想，

通过一次内部数据的泄漏，医院就会疏忽地将其他家庭成员的医疗状况跟你的家庭成员的信息搞混。毕竟，你们都居住在同一地址，可能性别相同，并且可能姓氏也一致。

虽然很容易想象，为该医疗保健服务支付的账单由于搞混，结果被浪费在了错误的人身上，但这种情况并不属于经济风险。法律上对IP地址严格对应具体个人方面的弱化要求，结合目前已经广泛使用的技术，便牵引出切实存在的风险的显现。风险就在于，即便利用高技术进行的医疗保健创新具有价值和利益，被开发的互联网技术与从中受益的人之间的现实关系也并未被充分体认。

“冒充别人”这种骗局有悠久的充满故事性的历史——无论在网上还是网下。但在上述场景当中，基于一个独立个体产生的数据便对应经由一个独立IP地址产生的数据这种假设之上，出于无心之失在一个亲属的医疗状况跟你自己的之间造成的混淆，将会在你一旦出现紧急情况进入急诊室抢救的时候带来灾难性的后果。

法官布朗的法律判决促动我们对隐私、身份、名誉以及数据所有权的各种细节的很有必要性的深入研究。在这个例子中，不能理解大数据技术与使用大数据技术的人之间关系的这种复杂性和细微之处，其直接影响将是切实产生极为严重的后果。

1.1 大数据为何“大”？

此刻你可能会问：“为什么不仅仅是一般的数据？”毕竟，一段时间以来，很多组织已经在大力寻找妥善处理数据的解决办法，不是吗？大数据作为时下流行的热词，其通俗定义基于传统数据库技术和前沿数据存储与处理技术及工具（诸如Hadoop集群、海量数据处理算法Bloom Filters及R语言数据分析工具等）之间的处理能力差异。大数据是指太“大”，以致无法使用诸如SQL等传统数据库协议进行分析和处理的数据（这种定义方式使大数据成为一个可能随着时间不断发展衍化的概念术语，即当下的所谓“大”数据可能十分迅速地变得“小”起来）。从这种认识角度出发，数据体量大小仅仅是这类新技术的一个方面而已。其风险与伦理考量也要从其他几个相关因素入手。

企业需要驾驭的信息的复杂性，由于可用信息之体量、种类和流转速率的变化而呈指数增长。这些因素所产生的问题，是人们在之前的商业经营过程中尚未遭遇过的。

当下数据的产生在体量上是惊人的。我们生活在一个全世界预期产生的数据量正以“艾字节”和“泽字节”进行计量的时代。到2025年，据预测互联网上的数据量将超过整个地球上全部人类的脑容量的总和。

此外，源数据类型的多样化趋势正在随着前沿技术的发展迅速加剧。车载监控反馈的各种性能指标、制造平台的现场测量数据、形形色色的医疗保健设备，以及越来越多的智能电

网能源设施，都在产生各种数据。

更重要的是，它们产生数据的速率很快。随着数据源数量的增加和数据格式的不断多样化发展，数据产生、获取、处理及输出的流转速率也同样呈指数增长。据广泛报道，全世界90%的数据是在最近两年中产生的 (<http://www.economist.com/node/21537967>)。大数据革命已然催生巨大变革，使处理复杂事件、抓取线上交互数据、开发移动计算产品与服务及实时处理很多大数据相关事件等方面的能力显著提升。

在近几年与利用大数据技术的组织一同研究的过程中，我们越发明显地发现，在变化多端的各种与大数据有关的情境中，对什么才是应该面对的伦理问题和应该采取的伦理制约，人们仍意见不一。在商业环境中进行大数据伦理的探讨，如果没有一个具有范式性与明晰性的框架，人们便常常会回归到他们各自的道德观点。尽管这不失为一个不错的起点，但这样很快便会陷入类似“但那实在太可怕了”“不，并不是如此”的争论状态，最终毫无进展。在此类案例中，最常发生的情况便是讨论因挫折而陷入僵局、会谈终止，而问题没有得到解决。而伦理无意识造成的潜在危害则能够快速压倒大数据创新带来的预期价值。

因此，当商业开拓者醉心于他们通过基于大数据技术之上的范围广泛的新产品和服务的设计与开发所能产生的可期利益时，可利用信息的体量、种类、速率却在引发新的问题。这些问题的一部分就是海量与人相关的各种属性、行为、偏好、关系、地点数据的获取、存储及利用所牵涉到的。

从根本上，这些问题属于伦理范畴。它们涉及价值观及该怎样运用大数据进行产品与服务创新。而价值观在创新主体如何对有益创新的承诺与有害风险进行平衡的问题上处于核心地位。无论你是否意识到你的价值观，它们都决定了你会怎样构想和实施主要基于萃集自海量数据的信息的产品与服务的设计方案。在权衡预期收益与意外风险时，它们是你进行衡量的关键性输入变量。

上述说明意味着对于大数据所能带来的创新，其风险与收益之间存在一个应该达到的平衡。本书的部分意图在于，帮助组织开发一种用以进行明晰有效的伦理检讨的框架，来帮助维持那个平衡点。

1.2 大数据“强加”了什么？

社会团体、政府部门乃至法制系统尚未适应这个正在到来的大数据时代的冲击，诸如透明性、关联性、聚合性等带来的问题。世界各国政府正在快速起草、讨论、出台新的法律法规。

不过一两代人之前，一个人能轻而易举地抛下与公民身份有关的一切瓜葛，在国境之内把

自己完全隐藏起来。而这在今天是不可能的，一个人的大多数行为都会产生数据踪迹，具备一定素质、装备简单的侦查人员便能沿着这些踪迹来到你面前 (http://www.wired.com/vanish/2009/11/ff_vanish2/)。

大数据具有永存性，并且它永存的形式是商业领域和社会领域之前从未经历过的。美国国会图书馆正在对2006年以来推特上的全部推文进行存档。而国会图书馆一旦把什么存档，就意味着他们希望将其永久保存。Facebook已将删除账号操作默认设置为并非彻底删除与账号相关联的全部数据 (<http://arstechnica.com/gadgets/2012/05/onfacebook-deleting-an-app-doesnt-delete-your-data-from-their-system/>)。

埃里克·弗里曼和大卫·格勒恩特尔创造了“生命之流”^{注3}这一说法：

一条以时间为序的连续存档记录，其功能相当于人们线上数字生活的日志。你自己创建的和别人发送给你的每一事件记录都被存储于你的生命之流中。生命之流的尾端则由以你最早接入网络时为始的过往记录构成。从尾端朝前浏览，你的生命之流越来越多地由更近切的记录填充——进行中的文本或新电子邮件。穿插其间其他形式数字记录（图片、通信、账单、视频、音频、软件）都被保存其中。跟随流的前端向未来移动，会发现流将由你需要的备忘录、日程表项、行动计划列表组成。

弗里曼和格勒恩特尔试图借“生命之流”预示处理个人线上数字信息的软件体系结构，此概念对理解大数据的永存性会如何影响个体生活的关键本质特征同样有用。大数据通常包含“元数据”，它能增加信息的层次（一层或几层），这些信息与我们每一个作为个体的人息息相关，对应着反映我们存在的种种客观现实情况。例如，大数据技术与体系结构能实现将你在推特上发推文时所在的现实地理位置关联到每一条推文。

并且这些附加的层级具有透明性。它们能够容纳海量的实质上游离于既有财务交易或社交互动记录的附加信息。大数据能重现你遍布全球的任何出行经历。大数据还能有目的地补充必要信息，将人个性的不同方面以我们有时无法完全控制的途径关联在一起。你在春假时玩乐的照片可能不经意间被当作相关材料在求职时受到关注，大数据在类似情况中已然显著改变了个人经营名誉的方式。

数据追踪仅仅是众多大数据技术相对以往实现对人类行为和活动更广泛、更深入的洞察介入之一例。各路开拓创新者已经认识到将这些洞察转化为新的有价值的产品和服务的潜能。数据的这一财富属性有望提升营销、管理、教育、研究、开发、保健、行政、服务以及我们日常生活方方面面的水平。大数据已经被应用于考察教育政策并对其进行有效的改

注3：<http://cs-www.cs.yale.edu/homes/freeman/lifestreams.html>